Contents

Preface *xi*
Roadmap to the Syllabus *xiii*

1. Probability **1.1–1.57**

1.1 Introduction *1.1*
1.2 Some Important Terms and Concepts *1.1*
1.3 Definitions of Probability *1.3*
1.4 Theorems on Probability *1.13*
1.5 Conditional Probability *1.25*
1.6 Multiplicative Theorem for Independent Events *1.25*
1.7 Bayes' Theorem *1.47*

20%

14Marks

2. Random Variables **2.1–2.83**

2.1 Introduction *2.1*
2.2 Random Variables *2.2*
2.3 Probability Mass Function *2.3*
2.4 Discrete Distribution Function *2.4*
2.5 Probability Density Function *2.18*
2.6 Continuous Distribution Function *2.18*
2.7 Two-Dimensional Discrete Random Variables *2.41*
2.8 Two-Dimensional Continuous Random Variables *2.56*

3. Basic Statistics **3.1–3.96**

3.1 Introduction *3.1*
3.2 Measures of Central Tendency *3.2*
3.3 Measures of Dispersion *3.3*
3.4 Moments *3.18*
3.5 Skewness *3.25*
3.6 Kurtosis *3.26*
3.7 Measures of Statistics for Continuous Random Variables *3.32*
3.8 Expected Values of Two Dimensional Random Variables *3.68*
3.9 Bounds on Probabilities *3.84*
3.10 Chebyshev's Inequality *3.84*

14 Marks

4. **Correlation and Regression** 4.1–4.56

20%.

4.1 Introduction *4.1*
4.2 Correlation *4.2*
4.3 Types of Correlations *4.2*
4.4 Methods of Studying Correlation *4.3*
4.5 Scatter Diagram *4.4*
4.6 Simple Graph *4.5*
4.7 Karl Pearson's Coefficient of Correlation *4.5*
4.8 Properties of Coefficient of Correlation *4.6*
4.9 Rank Correlation *4.22*
4.10 Regression *4.29*
4.11 Types of Regression *4.30*
4.12 Methods of Studying Regression *4.30*
4.13 Lines of Regression *4.31*
4.14 Regression Coefficients *4.31*
4.15 Properties of Regression Coefficients *4.34*
4.16 Properties of Lines of Regression (Linear Regression) *4.35*

5. **Some Special Probability Distributions** 5.1–5.104

25%

18 Marks

5.1 Introduction *5.1*
5.2 Binomial Distribution *5.2*
5.3 Poisson Distribution *5.27*
5.4 Normal Distribution *5.53*
5.5 Exponential Distribution *5.79*
5.6 Gamma Distribution *5.96*

6. **Applied Statistics: Test of Hypothesis** 6.1–6.86

25%

18 Marks

6.1 Introduction *6.1*
6.2 Terms Related to Tests of Hypothesis *6.2*
6.3 Procedure for Testing of Hypothesis *6.5*
6.4 Test of Significance for Large Samples *6.6*
6.5 Test of Significance for Single Proportion – Large Samples *6.8*
6.6 Test of Significance for Difference of Proportions – Large Samples *6.13*
6.7 Test of Significance for Single Mean – Large Samples *6.21*
6.8 Test of Significance for Difference of Means – Large Samples *6.26*
6.9 Test of Significance for Difference of Standard Deviations – Large Samples *6.31*
6.10 Small Sample Tests *6.36*
6.11 Student's *t*-distribution *6.36*
6.12 *t*-test: Test of Significance for Single Mean *6.37*
6.13 *t*-test: Test of Significance for Difference of Means *6.42*
6.14 *t*-test: Test of Significance for Correlation Coefficients *6.51*
6.15 Snedecor's *F*-test for Ratio of Variances *6.55*

6.16 Chi-square (χ^2) Test *6.65*
6.17 Chi-square Test: Goodness of Fit *6.66*
6.18 Chi-square Test for Independence of Attributes *6.74*

7. Curve Fitting 10%. (7 Marks) **7.1–7.26**

7.1 Introduction *7.1*
7.2 Least Square Method *7.2*
7.3 Fitting of Linear Curves *7.2*
7.4 Fitting of Quadratic Curves *7.10*
7.5 Fitting of Exponential and Logarithmic Curves *7.18*

Index ***I.1–I.4***

December

GTU. Winter 2019.

Chap=1, chap.2 → 14 Marks
Chap 3, chap4 → 14 Marks
Chap=5 → 18 Marks
Chap=6 → 17 Marks
Chap=7 → 7 Marks

70 Marks.

from:- D.G. BORAD

-: Shreenathji Engineering Zone :

D.Patel

CHAPTER 1

Probability

Chapter Outline

1.1 Introduction
1.2 Some Important Terms and Concepts
1.3 Definitions of Probability
1.4 Theorems on Probability
1.5 Conditional Probability
1.6 Multiplicative Theorem for Independent Events
1.7 Bayes' Theorem

1.1 INTRODUCTION

The concept of probability originated from the analysis of the games of chance. Even today, a large number of problems exist which are based on the games of chance, such as tossing of a coin, throwing of dice, and playing of cards. The utility of probability in business and economics is most emphatically revealed in the field of predictions for the future. Probability is a concept which measures the degree of uncertainty and that of certainty as a corollary.

The word *probability* or 'chance' is used commonly in day-to-day life. Daily, we come across the sentences like, 'it may rain today', 'India may win the forthcoming cricket match against Sri Lanka', 'the chances of making profits by investing in shares of Company A are very bright, etc. Each of the above sentences involves an element of uncertainty. A numerical measure of uncertainty is provided by a very important branch of mathematics called *theory of probability*. Before we study the probability theory in detail, it is appropriate to explain certain terms which are essential for the study of the theory of probability.

1.2 SOME IMPORTANT TERMS AND CONCEPTS

1. Random Experiment If an experiment is conducted, any number of times, under identical conditions, there is a set of all possible outcomes associated with it.

If the outcome is not unique but may be any one of the possible outcomes, the experiment is called a random experiment, e.g., tossing a coin, throwing a dice.

2. Outcome The result of a random experiment is called an outcome. For example, consider the following:

(a) Suppose a random experiment is 'a coin is tossed'. This experiment gives two possible outcomes—head or tail.

(b) Suppose a random experiment is 'a dice is thrown'. This experiment gives six possible outcomes—1, 2, 3, 4, 5 or 6—on the uppermost face of a dice.

3. Trial and Event Any particular performance of a random experiment is called a trial and outcome. A combination of outcomes is called an event. For example, consider the following:

(a) Tossing of a coin is a trial, and getting a head or tail is an event.

(b) Throwing of a dice is a trial and getting 1 or 2 or 3 or 4 or 5 or 6 is an event.

4. Exhaustive Event The total number of possible outcomes of a random experiment is called an exhaustive event. For example, consider the following:

(a) In tossing of a coin, there are two exhaustive events, viz., head and tail.

(b) In throwing of a dice, there are six exhaustive events, getting 1 or 2 or 3 or 4 or 5 or 6.

5. Mutually Exclusive Events Events are said to be mutually exclusive if the occurrence of one of them precludes the occurrence of all others in the same trial, i.e., they cannot occur simultaneously. For example, consider the following:

(a) In tossing a coin, the events head or tail are mutually exclusive since both head and tail cannot occur at the same time.

(b) In throwing a dice, all the six events, i.e., getting 1 or 2 or 3 or 4 or 5 or 6 are mutually exclusive events.

6. Equally Likely Events The outcomes of a random experiment are said to be equally likely if the occurrence of none of them is expected in preference to others. For example, consider the following:

(a) In tossing a coin, head or tail are equally likely events.

(b) In throwing a dice, all the six faces are equally likely events.

7. Independent Events Events are said to be independent if the occurrence of an event does not have any effect on the occurrence of other events. For example, consider the following:

(a) In tossing a coin, the event of getting a head in the first toss is independent of getting a head in the second, third, and subsequent tosses.

(b) In throwing a dice, the result of the first throw does not affect the result of the second throw.

8. Favourable Events The favourable events in a random experiment are the number of outcomes which entail the occurrence of the event. For example, consider the following:

In throwing of two dice, the favourable events of getting the sum 5 is (1, 4), (4, 1), (2, 3), (3, 2), i.e., 4.

1.3 DEFINITIONS OF PROBABILITY

1.3.1 Classical Definition of Probability

Let n be the number of equally likely, mutually exclusive, and exhaustive outcomes of a random experiment. Let m be number of the outcomes which are favourable to the occurrence of an event A. The probability of event A occurring, denoted by $P(A)$, is given by

$$P(A) = \frac{\text{Number of outcomes favourable to } A}{\text{Number of exhaustive outcomes}} = \frac{m}{n}$$

1.3.2 Empirical or Statistical Definition of Probability

If an experiment is repeated a large number of times under identical conditions, the limiting value of the ratio of the number of times the event A occurs to the total number of trials of the experiment as the number of trials increase indefinitely is called the probability of occurrence of the event A.

Let $P(A)$ be the probability of occurrence of the event A. Let m be the number of times in which an event A occurs in a series of n trials.

$$P(A) = \lim_{n\to\infty} \frac{m}{n}, \text{ provided the limit is finite and unique.}$$

1.3.3 Axiomatic Definition of Probability

Before discussing the axiomatic definition of probability, it is necessary to explain certain concepts that are necessary to its understanding.

1. Sample Space A set of all possible outcomes of a random experiment is called a sample space. Each element of the set is called a *sample point* or a *simple event* or an *elementary event*.

The sample space of a random experiment is denoted by S. For example, consider the following:

(a) In a random experiment of tossing of a coin, the sample space consists of two elementary events.

$$S = \{H, T\}$$

(b) In a random experiment of throwing of a dice, the sample space consists of six elementary events.

$$S = \{1, 2, 3, 4, 5, 6\}$$

The elements of S can either be single elements or ordered pairs. If two coins are tossed, each element of the sample space consists of the following ordered pairs:

$$S = \{(H, H), (H, T), (T, H), (T, T)\}$$

2. Event Any subset of a sample space is called an event. In the experiment of throwing of a dice, the sample space is $S = \{1, 2, 3, 4, 5, 6\}$. Let A be the event that an odd number appears on the dice. Then $A = \{1, 3, 5\}$ is a subset of S. Similarly, let B be the event of getting a number greater than 3. Then $B = \{4, 5, 6\}$ is another subset of S.

Definition of Probability Let S be a sample space of an experiment and A be any event of this sample space. The probability $P(A)$ of the event A is defined as the real-value set function which associates a real value corresponding to a subset A of the sample space S. The probability $P(A)$ satisfies the following three axioms.

Axiom I: $P(A) \geq 0$, i.e., the probability of an event is a nonnegative number.

Axiom II: $P(S) = 1$, i.e., the probability of an event that is certain to occur must be equal to unity.

Axiom III: If $A_1, A_2, \ldots, A_n$ are finite mutually exclusive events then

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$$

$$= \sum_{i=1}^{n} P(A_i)$$

i.e., the probability of a union of mutually exclusive events is the sum of probabilities of the events themselves.

Example 1

What is the probability that a leap year selected at random will have 53 Sundays?

Solution

A leap year has 366 days, i.e., 52 weeks and 2 days. These 2 days can occur in the following possible ways:

(i) Monday and Tuesday
(ii) Tuesday and Wednesday
(iii) Wednesday and Thursday
(iv) Thursday and Friday
(v) Friday and Saturday
(vi) Saturday and Sunday
(vii) Sunday and Monday

Number of exhaustive cases $n = 7$

Number of favourable cases $m = 2$

Let A be the event of getting 53 Sundays in a leap year.

$$P(A) = \frac{m}{n} = \frac{2}{7}$$

Example 2

Three unbiased coins are tossed. Find the probability of getting (i) exactly two heads, (ii) at least one tail, (iii) at most two heads, (iv) a head on the second coin, and (v) exactly two heads in succession.

Solution

When three coins are tossed, the sample space S is given by

$$S = \{HHH, HTH, THH, HHT, TTT, THT, TTH, HTT\}$$

$$n(s) = 8$$

(i) Let A be the event of getting exactly two heads.

$$A = \{HTH, THH, HHT\}$$

$$n(A) = 3$$

$$P(A) = \frac{n(A)}{n(S)} = \frac{3}{8}$$

(ii) Let B be the event of getting at least one tail.

$$B = \{HTH, THH, HHT, TTT, THT, TTH, HTT\}$$

$$n(B) = 7$$

$$P(B) = \frac{n(B)}{n(S)} = \frac{7}{8}$$

(iii) Let C be the event of getting at most two heads.

$$C = \{HTH, THH, HHT, TTT, THT, TTH, HTT\}$$

$$n(C) = 7$$

$$P(C) = \frac{n(C)}{n(S)} = \frac{7}{8}$$

(iv) Let D be the event of getting a head on the second coin.

$$D = \{HHH, THH, HHT, THT\}$$

$$n(D) = 4$$

$$P(D) = \frac{n(D)}{n(S)} = \frac{4}{8} = \frac{1}{2}$$

(v) Let E be the event of getting two heads in succession.

$$E = \{HH[illegible], THH, HHT\}$$

$$n(E) = 3$$

$$P(E) = \frac{n(E)}{n(S)} = \frac{3}{8}$$

Example 3

A fair dice is thrown. Find the probability of getting (i) an even number, (ii) a perfect square, and (iii) an integer greater than or equal to 3.

Solution

When a dice is thrown, the sample space S is given by

$$S = \{1, 2, 3, 4, 5, 6\}$$

$$n(S) = 6$$

(i) Let A be the event of getting an even number.

$$A = \{2, 4, 6\}$$

$$n(A) = 3$$

$$P(A) = \frac{n(A)}{n(S)} = \frac{3}{6} = \frac{1}{2}$$

(ii) Let B be the event of getting a perfect square.

$$B = \{1, 4\}$$

$$n(B) = 2$$

$$P(B) = \frac{n(B)}{n(S)} = \frac{2}{6} = \frac{1}{3}$$

(iii) Let C be the event of getting an integer greater than or equal to 3.

$$C = \{3, 4, 5, 6\}$$

$$n(C) = 4$$

$$P(C) = \frac{n(C)}{n(S)} = \frac{4}{6} = \frac{2}{3}$$

Example 4

A card is drawn from a well-shuffled pack of 52 cards. Find the probability of (i) getting a king card, (ii) getting a face card, (iii) getting a red card, (iv) getting a card between 2 and 7, both inclusive, and (v) getting a card between 2 and 8, both exclusive.

Solution

Total number of cards = 52

One card out of 52 cards can be drawn in ways.

$$n(S) = {}^{52}C_1 = 52$$

(i) Let A be the event of getting a king card. There are 4 king cards and one of them can be drawn in ${}^{4}C_1$ ways.

$$n(A) = {}^{4}C_1 = 4$$

$$P(A) = \frac{n(A)}{n(S)} = \frac{4}{52} = \frac{1}{13}$$

(ii) Let B be the event of getting a face card. There are 12 face cards and one of them can be drawn in ${}^{12}C_1$ ways.

$$n(B) = {}^{12}C_1 = 12$$

$$P(B) = \frac{n(B)}{n(S)} = \frac{12}{52} = \frac{3}{13}$$

(iii) Let C be the event of getting a red card. There are 26 red cards and one of them can be drawn in ${}^{26}C_1$ ways.

$$n(C) = {}^{26}C_1 = 26$$

$$P(C) = \frac{n(C)}{n(S)} = \frac{26}{52} = \frac{1}{2}$$

(iv) Let D be the event of getting a card between 2 and 7, both inclusive. There are 6 such cards in each suit giving a total of $6 \times 4 = 24$ cards. One of them can be drawn in ${}^{24}C_1$ ways.

$$n(D) = {}^{24}C_1 = 24$$

$$P(D) = \frac{n(D)}{n(S)} = \frac{24}{52} = \frac{6}{13}$$

(v) Let E be the event of getting a card between 2 and 8, both exclusive. There are 5 such cards in each suit giving a total of $5 \times 4 = 20$ cards. One of them can be drawn in ${}^{20}C_1$ ways.

$$n(E) = {}^{20}C_1 = 20$$

$$= \frac{n(E)}{n(S)} = \frac{20}{52} = \frac{5}{13}$$

Example 5

A bag contains 2 black, 3 red, and 5 blue balls. Three balls are drawn at random. Find the probability that the three balls drawn (i) are blue (ii) consist of 2 blue and 1 red ball, and (iii) consist of exactly one black ball.

Solution

Total number of balls = 10

3 balls out of 10 balls can be drawn in ${}^{10}C_3$ ways.

$$n(S) = {}^{10}C_3 = 120$$

(i) Let A be the event that the three balls drawn are blue. 3 blue balls out of 5 blue balls can be drawn in ${}^{5}C_3$ ways.

$$n(A) = {}^{5}C_3 = 10$$

$$P(A) = \frac{n(A)}{n(S)} = \frac{10}{120} = \frac{1}{12}$$

(ii) Let B be the event that the three balls drawn consist of 2 blue and 1 red ball. 2 blue balls out of 5 blue balls can be drawn in ${}^{5}C_2$ ways. 1 red ball out of 3 red balls can be drawn in ${}^{3}C_1$ ways.

$$n(B) = {}^{5}C_2 \times {}^{3}C_1 = 30$$

$$P(B) = \frac{n(B)}{n(S)} = \frac{30}{120} = \frac{1}{4}$$

(iii) Let C be the event that three balls drawn consist of exactly one black ball, i.e., remaining two balls can be drawn from 3 red and 5 blue balls. One black ball can be drawn from 2 black balls in ${}^{2}C_1$ ways and the remaining 2 balls can be drawn from 8 balls in ${}^{8}C_2$ ways.

$$n(C) = {}^{2}C_1 \times {}^{8}C_2 = 56$$

$$P(C) = \frac{n(C)}{n(S)} = \frac{56}{120} = \frac{7}{15}$$

Example 6

A class consists of 6 girls and 10 boys. If a committee of three is chosen at random from the class, find the probability that (i) three boys are selected, and (ii) exactly two girls are selected.

Solution

Total number of students = 16

A committee of 3 students from 16 students can be selected in ${}^{16}C_3$ ways.

$$n(S) = {}^{16}C_3 = 560$$

(i) Let A be the event that 3 boys are selected.

$$n(A) = {}^{10}C_3 = 120$$

$$P(A) = \frac{n(A)}{n(S)} = \frac{120}{560} = \frac{3}{14}$$

(ii) Let B be the event that exactly 2 girls are selected. 2 girls from 6 girls can be selected in ${}^{6}C_2$ ways and one boy from 10 boys can be selected in ${}^{10}C_1$ ways.

$$n(B) = {}^{6}C_2 \times {}^{10}C_1 = 150$$

$$P(B) = \frac{n(B)}{n(S)} = \frac{150}{560} = \frac{15}{16}$$

Example 7

From a collection of 10 bulbs, of which 4 are defective, 3 bulbs are selected at random and fitted into lamps. Find the probability that (i) all three bulbs glow, and (ii) the room is lit.

Solution

Total number of bulbs = 10

3 bulbs can be selected from 10 bulbs in ${}^{10}C_3$ ways.

$$n(S) = {}^{10}C_3 = 120$$

(i) Let A be event that all three bulbs glow. This event will occur when 3 bulbs are selected from 6 nondefective bulbs in ${}^{6}C_3$ ways.

$$n(A) = {}^{6}C_3 = 20$$

$$P(A) = \frac{n(A)}{n(S)} = \frac{20}{120} = \frac{1}{6}$$

(ii) Let B be the event that the room is lit. Let $\bar{B}$ be the event that the room is dark. The event $\bar{B}$ will occur when 3 bulbs are selected from 4 defective bulbs in ${}^{4}C_3$ ways.

$$n(\bar{B}) = {}^{4}C_3 = 4$$

$$P(\bar{B}) = \frac{n(\bar{B})}{n(S)} = \frac{4}{120} = \frac{1}{30}$$

$$\therefore \quad P(B) = 1 - P(\bar{B}) = 1 - \frac{1}{30} = \frac{29}{30}$$

Example 8

There are 20 tickets numbered 1, 2, ..., 20. One ticket is drawn at random. Find the probability that the ticket bears a number which is (i) even, (ii) a perfect square, and (iii) multiple of 3.

Solution

There are 20 tickets numbered from 1 to 20.

$n(S) = 20$

(i) Let A be the event that a ticket bears a number which is even.

$$A = \{2, 4, 6, 8, 10, 12, 14, 16, 18, 20\}$$
$$n(A) = 10$$
$$P(A) = \frac{n(A)}{n(S)} = \frac{10}{20} = \frac{1}{2}$$

(ii) Let B the event that a ticket bears a number which is a perfect square.

$$B = \{1, 4, 9, 16\}$$
$$n(B) = 4$$
$$P(B) = \frac{n(B)}{n(S)} = \frac{4}{20} = \frac{1}{5}$$

(iii) Let C be the event that a ticket bears a number which is a multiple of 3.

$$C = \{3, 6, 9, 12, 15, 18\}$$
$$n(C) = 6$$
$$P(C) = \frac{n(C)}{n(S)} = \frac{6}{20} = \frac{3}{10}$$

Example 9

Four letters of the word 'THURSDAY' are arranged in all possible ways. Find the probability that the word formed is 'HURT'.

Solution

Total number of letters in the word 'THURSDAY' = 8

Four letters from 8 letters can be arranged in ${}^8P_4$ ways.

$$n(S) = {}^8P_4 = 1680$$

Let A be the event that the word formed is 'HURT'. The word 'HURT' can be formed in one way only.

$$n(A) = 1$$
$$P(A) = \frac{n(A)}{n(S)} = \frac{1}{1680}$$

Example 10

A bag contains 5 red, 4 blue, and m green balls. If the probability of getting two green balls when two balls are selected at random is $\frac{1}{7}$, *find m.*

Solution

Total number of balls = $5 + 4 + m = 9 + m$

2 balls out of $9 + m$ balls can be drawn in ${}^{9+m}C_2$ ways.

$$n(S) = {}^{9+m}C_2$$

Let A be the event that both the balls drawn are green.

2 green balls out of m green balls can be drawn in ${}^mC_2$ ways.

$$n(A) = {}^mC_2$$
$$P(A) = \frac{n(A)}{n(S)} = \frac{{}^mC_2}{{}^{9+m}C_2}$$

But $P(A) = \frac{1}{7}$

$$\frac{{}^mC_2}{{}^{9+m}C_2} = \frac{1}{7}$$
$$\frac{m(m-1)}{(m+9)(m+8)} = \frac{1}{7}$$
$$(m+9)(m+8) = 7m(m-1)$$
$$m^2 + 17m + 72 = 7m^2 - 7m$$
$$6m^2 - 24m - 72 = 0$$
$$3m^2 - 12m - 36 = 0$$
$$3m^2 - 18m + 6m - 36 = 0$$
$$3m(m-6) + 6(m-6) = 0$$
$$(3m+6)(m-6) = 0$$
$$3m + 6 = 0 \quad \text{or} \quad m - 6 = 0$$
$$m = -2 \quad \text{or} \quad m = 6$$

But $m \neq -2$

$\therefore$ $m = 6$

EXERCISE 1.1

1. A card is drawn at random from a pack of 52 cards. Find the probability that the card drawn is (i) an ace card, and (ii) a club card.

$\left[\text{Ans.: (i) } \frac{1}{13} \text{ (ii) } \frac{1}{4}\right]$

2. An unbiased coin is tossed twice. Find the probability of (i) exactly one head, (ii) at most one head, (iii) at least one head, and (iv) same face on both the coins.

$$\left[\textbf{Ans.:} \text{ (i) } \frac{1}{2} \text{ (ii) } \frac{3}{4} \text{ (iii) } \frac{3}{4} \text{ (iv) } \frac{1}{2}\right]$$

3. A fair dice is thrown thrice. Find the probability that the sum of the numbers obtained is 10.

$$\left[\textbf{Ans.:} \frac{1}{8}\right]$$

4. A ball is drawn at random from a box containing 12 red, 18 white, 19 blue, and 15 orange balls. Find the probability that (i) it is red or blue, and (ii) it is white, blue, or orange.

$$\left[\textbf{Ans.:} \text{ (i) } \frac{2}{5} \text{ (ii) } \frac{43}{55}\right]$$

5. Eight boys and three girls are to sit in a row for a photograph. Find the probability that no two girls are together.

$$\left[\textbf{Ans.:} \frac{28}{55}\right]$$

6. If four persons are chosen from a group of 3 men, 2 women, and 4 children, find the probability that exactly two of them will be children.

$$\left[\textbf{Ans.:} \frac{10}{21}\right]$$

7. A box contains 2 white, 3 red, and 5 black balls. Three balls are drawn at random. What is the probability that they will be of different colours?

$$\left[\textbf{Ans.:} \frac{1}{4}\right]$$

8. Two cards are drawn from a well-shuffled pack of 52 cards. Find the probability of getting (i) 2 king cards, (ii) 1 king card and 1 queen card, and (iii) 1 king card and 1 spade card.

$$\left[\textbf{Ans.:} \text{ (i) } \frac{1}{221} \text{ (ii) } \frac{8}{663} \text{ (iii) } \frac{1}{26}\right]$$

9. A four-digit number is to be formed using the digits 0, 1, 2, 3, 4, 5. All the digits are to be different. Find the probability that the digit formed is (i) odd, (ii) greater than 4000, (iii) greater than 3400, and (iv) a multiple of 5.

$$\left[\textbf{Ans.:} \text{ (i) } \frac{12}{25} \text{ (ii) } \frac{2}{5} \text{ (iii) } \frac{12}{25} \text{ (iv) } \frac{9}{25}\right]$$

10. 3 books of physics, 4 books of chemistry, and 5 books of mathematics are arranged in a shelf. Find the probability that (i) no physics books are together, (ii) chemistry books are always together, and (iii) books of the same subjects are together.

$$\left[\textbf{Ans.:} \text{ (i) } \frac{6}{11} \text{ (ii) } \frac{1}{55} \text{ (iii) } \frac{1}{4620}\right]$$

11. 8 boys and 2 girls are to be seated at random in a row for a photograph. Find the probability that (i) the girls sit together, and (ii) the girls occupy 3rd and 7th seats.

$$\left[\textbf{Ans.:} \text{ (i) } \frac{1}{5} \text{ (ii) } \frac{1}{45}\right]$$

12. A committee of 4 is to be formed from 15 boys and 3 girls. Find the probability that the committee contains (i) 2 boys and 2 girls, (ii) exactly one girl, (iii) one particular girl, and (iv) two particular girls.

$$\left[\textbf{Ans.:} \text{ (i) } \frac{7}{68} \text{ (ii) } \frac{91}{204} \text{ (iii) } \frac{2}{9} \text{ (iv) } \frac{2}{51}\right]$$

13. If the letters of the word REGULATIONS are arranged at random, what is the probability that there will be exactly four letters between R and E?

$$\left[\textbf{Ans.:} \frac{6}{55}\right]$$

14. Find the probability that there will be 5 Sundays in the month of October.

$$\left[\textbf{Ans.:} \frac{3}{7}\right]$$

1.4 THEOREMS ON PROBABILITY

Theorem 1 The probability of an impossible event is zero, i.e., $P(\phi) = 0$, where ϕ is a null set.

Proof An event which has no sample points is called an impossible event and is denoted by ϕ.

For a sample space S of an experiment,

$$S \cup \phi = S$$

Taking probability of both the sides,

$$P(S \cup \phi) = P(S)$$

Since S and ϕ are mutually exclusive events,

$$P(S) + P(\phi) = P(S) \qquad \text{[Using Axiom III]}$$

$$\therefore \quad P(\phi) = 0$$

Theorem 2 The probability of the complementary event $\bar{A}$ of A is

$$P(\bar{A}) = 1 - P(A)$$

Proof Let A be an event in the sample space S.

$$A \cup \bar{A} = S$$

$$P(A \cup \bar{A}) = P(S)$$

Since A and $\bar{A}$ are mutually exclusive events,

$$P(A) + P(\bar{A}) = P(S)$$

$$P(A) + P(\bar{A}) = 1 \qquad [\because \; P(S) = 1]$$

$$\therefore \quad P(\bar{A}) = 1 - P(A)$$

Note Since A and $\bar{A}$ are mutually exclusive events,

$$A \cup \bar{A} = S \text{ and } A \cap \bar{A} = \phi$$

Corollary Probability of an event is always less than or equal to one, i.e., $P(A) \le 1$

Proof $P(A) = 1 - P(\bar{A})$

$$P(A) \le 1 \qquad [\because \; P(\bar{A}) \ge 0 \text{ by Axiom I}]$$

De Morgan's Laws Since an event is a subset of a sample space, De Morgan's laws are applicable to events.

$$P(\overline{A \cup B}) = P(\bar{A} \cap \bar{B})$$

$$P(\overline{A \cap B}) = P(\bar{A} \cup \bar{B})$$

Theorem 3 For any two events A and B in a sample space S,

$$P(\bar{A} \cap B) = P(B) - P(A \cap B)$$

Proof From the Venn diagram (Fig. 1.1),

$$B = (A \cap B) \cup (\bar{A} \cap B)$$

$$P(B) = P\left[(A \cap B) \cup (\bar{A} \cap B)\right]$$

Since $(A \cap B)$ and $(\bar{A} \cap B)$ are mutually exclusive events,

$$P(B) = P(A \cap B) + P(\bar{A} \cap B)$$

$$P(\bar{A} \cap B) = P(B) - P(A \cap B)$$

Similarly, it can be shown that

$$P(A \cap \bar{B}) = P(A) - P(A \cap B)$$

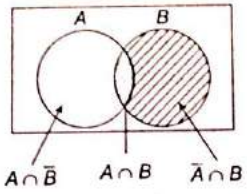


Fig. 1.1

Theorem 4 Additive Law of Probability (Addition Theorem)

The probability that at least one of the events A and B will occur is given by

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Proof From the Venn diagram (Fig. 1.1),

$$A \cup B = A \cup (\bar{A} \cap B)$$

$$P(A \cup B) = P\left[A \cup (\bar{A} \cap B)\right]$$

Since A and $(\bar{A} \cap B)$ are mutually exclusive events,

$$P(A \cup B) = P(A) + P(\bar{A} \cap B) \qquad \text{[Using Axiom III]}$$

$$= P(A) + P(B) - P(A \cap B) \qquad \text{[Using Theorem 3]}$$

Remarks

1. If A and B are mutually exclusive events, i.e., $A \cap B = \phi$ then $P(A \cap B) = 0$ according to Theorem 1.

   Hence, $P(A \cup B) = P(A) + P(B)$

2. The event $A \cup B$ (i.e., A or B) denotes the occurrence of either A or B or both. Alternately, it implies the occurrence of at least one of the two events.

   $A \cup B = A + B$

3. The event $A \cap B$ (i.e., A and B) is a compound or joint event that denotes the simultaneous occurrence of the two events.

   $A \cap B = AB$

Corollary 1 From the Venn diagram (Fig. 1.1),

$$P(A \cup B) = 1 - P(\bar{A} \cap \bar{B})$$

where $P(\bar{A} \cap \bar{B})$ is the probability that none of the events A and B occur simultaneously.

Corollary 2 $P(\text{Exactly one of } A \text{ and } B \text{ occurs}) = P\left[(A \cap \bar{B}) \cup (\bar{A} \cap B)\right]$

$$= P(A \cap \bar{B}) + P(\bar{A} \cap B) \qquad \left[\because \; (A \cap \bar{B}) \cap (\bar{A} \cap B) = \phi\right]$$

$$= P(A) - P(A \cap B) + P(B) - P(A \cap B) \qquad \text{[Using Theorem 3]}$$

$$= P(A) + P(B) - 2P(A \cap B)$$

$$= P(A \cup B) - P(A \cap B) \qquad \text{[Using Theorem 4]}$$

$= P$ (at least one of the two events occur)

$\quad - P$ (the two events occur simultaneously)

Corollary 3 The addition theorem can be applied for more than two events. If A, B, and C are three events of a sample space S then the probability of occurrence of at least one of them is given by.

$$\begin{aligned}P(A\cup B\cup C) &= P[A\cup(B\cup C)]\\ &= P(A)+P(B\cup C)-P[A\cap(B\cup C)]\\ &= P(A)+P(B\cup C)-P[A\cap B)\cup(A\cap C)]\\ &= P(A)+P(B)+P(C)-P(B\cap C)-P(A\cap B)-P(A\cap C)+P(A\cap B\cap C)\end{aligned}$$

[Applying Theorem 4 on second and third term]

Alternately, the probability of occurrence of at least one of the three events can also be written as

$$P(A\cup B\cup C)=1-P(\bar{A}\cap\bar{B}\cap\bar{C})$$

If A, B, and C are mutually exclusive events,

$$P(A\cup B\cup C)=P(A)+P(B)+P(C)$$

Corollary 4 The probability of occurrence of at least two of the three events is given by

$$\begin{aligned}P[A\cap B)\cup(B\cap C)\cup(A\cap C)] &= P(A\cap B)+P(B\cap C)+P(A\cap C)-3P(A\cap B\cap C)\\ &\quad +P(A\cap B\cap C) \qquad \text{[Using Corollary 3]}\\ &= P(A\cap B)+P(B\cap C)+P(A\cap C)-2P(A\cap B\cap C)\end{aligned}$$

Corollary 5 The probability of occurrence of exactly two of the three events is given by

$$\begin{aligned}&P[A\cap B\cap\bar{C})\cup(A\cap\bar{B}\cap C)\cup(\bar{A}\cap B\cap C)]\\ &= P[(A\cap B)\cup(B\cap C)\cup(A\cap C)]-P(A\cap B\cap C) \quad \text{[Using Corollary 2]}\\ &= P(A\cap B)+P(B\cap C)+P(A\cap C)-3P(A\cap B\cap C) \quad \text{[Using Corollary 4]}\end{aligned}$$

Corollary 6 The probability of occurrence of exactly one of the three events is given by

$$P[(A\cap\bar{B}\cap\bar{C})\cup(\bar{A}\cap B\cap\bar{C})\cup(\bar{A}\cap\bar{B}\cap C)]$$

$= P$(at least one of the three event occur) – P (at least two of the three events occur)

$$= P(A)+P(B)+P(C)-2P(A\cap B)-2P(B\cap C)-2P(A\cap C)+3P(A\cap B\cap C)$$

Example 1

A card is drawn from a well-shuffled pack of cards. What is the probability that it is either a spade or an ace?

Solution

Let A and B be the events of getting a spade and an ace card respectively.

$$P(A)=\frac{{}^{13}C_1}{{}^{52}C_1}=\frac{13}{52}$$

$$P(B)=\frac{{}^{4}C_1}{{}^{52}C_1}=\frac{4}{52}$$

$$P(A\cap B)=\frac{{}^{1}C_1}{{}^{52}C_1}=\frac{1}{52}$$

Probability of getting either a spade or an ace card

$$\begin{aligned}P(A\cup B) &= P(A)+P(B)-P(A\cap B)\\ &= \frac{13}{52}+\frac{4}{52}-\frac{1}{52}\\ &= \frac{4}{13}\end{aligned}$$

Example 2

Two cards are drawn from a pack of cards. Find the probability that they will be both red or both pictures.

Solution

Let A and B be the events that both cards drawn are red and pictures respectively.

$$P(A)=\frac{{}^{26}C_2}{{}^{52}C_2}=\frac{325}{1326}$$

$$P(B)=\frac{{}^{12}C_2}{{}^{52}C_2}=\frac{66}{1326}$$

$$P(A\cap B)=\frac{{}^{6}C_2}{{}^{52}C_2}=\frac{15}{1326}$$

Probability that both cards drawn are red or pictures

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)$$

$$= \frac{325}{1326} + \frac{66}{1326} - \frac{15}{1326}$$

$$= \frac{188}{663}$$

Example 3

The probability that a contractor will get a plumbing contract is $\frac{2}{3}$ *and the probability that he will not get an electric contract is* $\frac{5}{9}$. *If the probability of getting any one contract is* $\frac{4}{5}$, *what is the probability that he will get both the contracts?*

Solution

Let A and B be the events that the contractor will get plumbing and electric contracts respectively.

$$P(A) = \frac{2}{3},\quad P(\overline{B}) = \frac{5}{9},\quad P(A \cup B) = \frac{4}{5}$$

$$P(B) = 1 - P(\overline{B}) = 1 - \frac{5}{9} = \frac{4}{9}$$

Probability that the contractor will get any one contract

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Probability that the contractor will get both the contracts

$$P(A \cap B) = P(A) + P(B) - P(A \cup B)$$

$$= \frac{2}{3} + \frac{4}{9} - \frac{4}{5}$$

$$= \frac{14}{45}$$

Example 4

A person applies for a job in two firms A and B, the probability of his being selected in the firm A is 0.7 and being rejected in the firm B is 0.5. The probability of at least one of the applications being rejected is 0.6. What is the probability that he will be selected in one of the two firms?

Solution

Let A and B be the events that the person is selected in firms A and B respectively.

$$P(A) = 0.7,\quad P(\overline{B}) = 0.5,\quad P(\overline{A} \cup \overline{B}) = 0.6$$

$$P(\overline{A}) = 1 - P(A) = 1 - 0.7 = 0.3$$

$$P(B) = 1 - P(\overline{B}) = 1 - 0.5 = 0.5$$

$$P(\overline{A} \cup \overline{B}) = P(\overline{A}) + P(\overline{B}) - P(\overline{A} \cap \overline{B}) \quad \text{...(1)}$$

Probability that the person will be selected in one of the two firms

$$P(A \cup B) = 1 - P(\overline{A} \cap \overline{B})$$

$$= 1 - \left[P(\overline{A}) + P(\overline{B}) - P(\overline{A} \cup \overline{B})\right] \quad \text{[Using Eq. (1)]}$$

$$= 1 - (0.3 + 0.5 - 0.6)$$

$$= 0.8$$

Example 5

In a group of 1000 persons, there are 650 who can speak Hindi, 400 can speak English, and 150 can speak both Hindi and English. If a person is selected at random, what is the probability that he speaks (i) Hindi only, (ii) English only, (iii) only of the two languages, and (iv) at least one of the two languages?

Solution

Let A and B be the events that a person selected at random speaks Hindi and English respectively.

$$P(A) = \frac{650}{1000},\quad P(B) = \frac{400}{1000}, P(A \cap B) = \frac{150}{1000}$$

(i) Probability that a person selected at random speaks Hindi only

$$P(A \cap \overline{B}) = P(A) - P(A \cap B)$$

$$= \frac{650}{1000} - \frac{150}{1000}$$

$$= \frac{1}{2}$$

(ii) Probability that a person selected at random speaks English only

$$P(\overline{A} \cap B) = P(B) - P(A \cap B)$$

$$= \frac{400}{1000} - \frac{150}{1000}$$

$$= \frac{1}{4}$$

(iii) Probability that a person selected at random speaks only one of the languages.

$$P\left[(A\cap\bar{B})\cup(\bar{A}\cap B)\right]=P(A)+P(B)-2P(A\cap B)$$

$$=\frac{650}{1000}+\frac{400}{1000}-2\left(\frac{150}{1000}\right)$$

$$=\frac{3}{4}$$

(iv) Probability that a person selected at random speaks at least one of the two languages

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)$$

$$=\frac{650}{1000}+\frac{400}{1000}-\frac{150}{1000}$$

$$=\frac{9}{10}$$

Example 6

A box contains 4 white, 6 red, 5 black balls, and 5 balls of other colours. Two balls are drawn from the box at random. Find the probability that (i) both are white or both are red, and (ii) both are red or both are black.

Solution

Let A, B, and C be the events of drawing white, red and black balls from the box respectively.

$$P(A)=\frac{^4C_2}{^{20}C_2}=\frac{3}{95}$$

$$P(B)=\frac{^6C_2}{^{20}C_2}=\frac{3}{38}$$

$$P(C)=\frac{^5C_2}{^{20}C_2}=\frac{1}{19}$$

(i) Probability that the both balls are white or both are red

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)$$

$$=\frac{3}{95}+\frac{3}{38}-0$$

$$=\frac{21}{190}$$

(ii) Probability that both balls are red or both are black

$$P(B\cup C)=P(B)+P(C)-P(B\cap C)$$

$$=\frac{3}{38}+\frac{1}{19}-0$$

$$=\frac{5}{38}$$

Example 7

Three students A, B, C are in a running race. A and B have the same probability of winning and each is twice as likely to win as C. Find the probability that B or C wins.

Solution

Let A, B, and C be the events that students A, B, and C win the race respectively.

$$P(A)=P(B)=2P(C)$$

$$P(A)+P(B)+P(C)=1$$

$$2P(C)+2P(C)+P(C)=1$$

$$P(C)=\frac{1}{5}$$

$$\therefore \quad P(A)=\frac{2}{5}\text{ and }P(B)=\frac{2}{5}$$

Probability that student B or C wins

$$P(B\cup C)=P(B)+P(C)-P(B\cap C)$$

$$=\frac{2}{5}+\frac{1}{5}-0$$

$$=\frac{3}{5}$$

Example 8

A card is drawn from a pack of 52 cards. Find the probability of getting a king or a heart or a red card.

Solution

Let A, B and C be the events that the card drawn is a king, a heart and a red card respectively.

$$P(A)=\frac{{}^{4}C_1}{{}^{52}C_1}=\frac{4}{52}$$

$$P(B)=\frac{{}^{13}C_1}{{}^{52}C_1}=\frac{13}{52}$$

$$P(C)=\frac{{}^{26}C_1}{{}^{52}C_1}=\frac{26}{52}$$

$$P(A\cap B)=\frac{{}^{1}C_1}{{}^{52}C_1}=\frac{1}{52}$$

$$P(B\cap C)=\frac{{}^{13}C_1}{{}^{52}C_1}=\frac{13}{52}$$

$$P(A\cap C)=\frac{{}^{2}C_1}{{}^{52}C_1}=\frac{2}{52}$$

$$P(A\cap B\cap C)=\frac{{}^{1}C_1}{{}^{52}C_1}=\frac{1}{52}$$

Probability that the card drawn is a king or a heart or a red card.

$$\begin{aligned}P(A\cup B\cup C)&=P(A)+P(B)+P(C)-P(A\cap B)-P(B\cap C)-P(A\cap C)\\&\quad+P(A\cap B\cap C)\\&=\frac{4}{52}+\frac{13}{52}+\frac{26}{52}-\frac{1}{52}-\frac{13}{52}-\frac{2}{52}+\frac{1}{52}\\&=\frac{7}{3}\end{aligned}$$

Example 9

From a city, 3 newspapers A, B, C are being published. A is read by 20%, B is read by 16%, C is read by 14%, both A and B are read by 8%, both A and C are read by 5%, both B and C are read by 4% and all three A, B, C are read by 2%. What is the probability that a randomly chosen person (i) reads at least one of these newspapers, and (ii) reads one of these newspapers?

Solution

Let A, B, and C be the events that the person reads newspapers A, B, and C respectively.

$$P(A)=0.2,\qquad P(B)=0.16\qquad P(C)=0.14$$

$$P(A\cap B)=0.08,\qquad P(A\cap B)=0.05,\qquad P(B\cap C)=0.04$$

$$P(A\cap B\cap C)=0.02$$

(i) Probability that the person reads at least one of these newspapers

$$\begin{aligned}P(A\cup B\cup C)&=P(A)+P(B)+P(C)-P(A\cap B)-P(A\cap C)-P(B\cap C)\\&\quad+P(A\cap B\cap C)\\&=0.2+0.16+0.14-0.08-0.05-0.04+0.02\\&=0.35\end{aligned}$$

(ii) Probability that the person reads none of these newspapers

$$\begin{aligned}P(\bar{A}\cap\bar{B}\cap\bar{C})&=1-P(A\cup B\cup C)\\&=1-0.35\\&=0.65\end{aligned}$$

Alternatively, the problem can be solved by a Venn diagram (Fig. 1.2).

(i) $P(\text{the person reads at least one paper})=1-\frac{65}{100}=0.35$

(ii) $P(\text{the person reads none of these papers})=0.65$

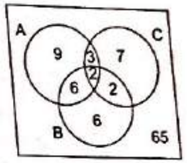


Fig. 1.2

EXERCISE 1.2

1. The probability that a student passes a Physics test is $\frac{2}{3}$ and the probability that he passes both Physics and English tests is $\frac{14}{45}$. The probability that he passes at least one test is $\frac{4}{5}$. What is the probability that the student passes the English test?

$\left[\text{Ans.: }\frac{4}{9}\right]$

2. What is the probability of drawing a black card or a king from a well-shuffled pack of playing cards?

$\left[\text{Ans.: }\frac{7}{13}\right]$

3. A pair of unbiased dice is thrown. Find the probability that (i) the sum of spots is either 5 or 10, and (ii) either there is a doublet or a sum less than 6.

$\left[\text{Ans.: (i) }\frac{7}{36}\text{ (ii) }\frac{7}{18}\right]$

4. From a pack of well-shuffled cards, a card is drawn at random. What is the probability that the card drawn is a diamond card or a king card?

$\left[\text{Ans.: } \frac{4}{13}\right]$

5. A bag contains 6 red, 5 blue, 3 white, and 4 black balls. A ball is drawn at random. Find the probability that the ball is (i) red or black, and (ii) neither red or black.

$\left[\text{Ans.: (i) } \frac{5}{9} \text{ (ii) } \frac{4}{9}\right]$

6. There are 100 lottery tickets, numbered from 1 to 100. One of them is drawn at random. What is the probability that the number on it is a multiple of 5 or 7?

$\left[\text{Ans.: } \frac{8}{25}\right]$

7. From a group of 6 boys and 4 girls, a committee of 3 is to be formed. Find the probability that the committee will include (i) all three boys or all three girls, (ii) at most two girls, and (iii) at least one girl.

$\left[\text{Ans.: (i) } \frac{1}{5} \text{ (ii) } \frac{29}{30} \text{ (iii) } \frac{5}{6}\right]$

8. From a pack of 52 cards, three cards are drawn at random. Find the probability that (i) all three will be aces or all three kings, (ii) all three are pictures or all three are aces, (iii) none is a picture, (iv) at least one is a picture, (v) none is a spade, (vi) at most two are spades, and (vii) at least one is a spade.

$$\left[\begin{array}{l}\text{Ans.: (i) } \frac{2}{5225} \text{ (ii) } \frac{56}{5225} \text{ (iii) } \frac{38}{85} \text{ (iv) } \frac{47}{85} \\ \text{(v) } \frac{703}{1700} \text{ (vi) } \frac{839}{850} \text{ (vii) } \frac{997}{1700}\end{array}\right]$$

9. From a set of 16 cards numbered 1 to 16, one card is drawn at random. Find the probability that (i) the number obtained is divisible by 3 or 7, and (ii) not divisible by 3 and 7.

$\left[\text{Ans.: (i) } \frac{7}{16} \text{ (ii) } \frac{9}{16}\right]$

10. There are 12 bulbs in a basket of which 4 are working. A person tries to fit them in 3 sockets choosing 3 of the bulbs at random. What is the probability that there will be (i) some light, and (ii) no light in the room?

$\left[\text{Ans.: (i) } \frac{41}{55} \text{ (ii) } \frac{14}{55}\right]$

1.5 CONDITIONAL PROBABILITY

For any two events A and B in a sample space S, the probability of their simultaneous occurrence, i.e., both the events occurrings simultaneously is given by

$$P(A \cap B) = P(A)\, P(B/A)$$

or

$$P(A \cap B) = P(B)\, P(A/B)$$

where $P(B/A)$ is the conditional probability of B given that A has already occurred. $P(A/B)$ is the conditional probability of A given that B has already occurred.

1.6 MULTIPLICATIVE THEOREM FOR INDEPENDENT EVENTS

If A and B are two independent events, the probability of their simultaneous occurrence is given by

$$P(A \cap B) = P(A)\, P(B)$$

$$P(A \cap B) = P(B)\, P(A/B) \qquad \text{...(1.1)}$$

Proof $A = (A \cap B) \cup (A \cap \bar{B})$

Since $(A \cap B)$ and $(A \cap \bar{B})$ are mutually exclusive events,

$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap \bar{B}) \qquad \text{[Using Axiom III]} \\ &= P(B)\, P(A/B) + P(\bar{B})\, P(A/\bar{B}) \end{aligned}$$

If A and B are independent events, the proportion of A's in B is equal to proportion of A's in $\bar{B}$, i.e., $P(A/B) = P(A/\bar{B})$.

$$\begin{aligned} P(A) &= P(A/B)\left[P(B) + P(\bar{B})\right] \\ &= P(A/B) \end{aligned}$$

Substituting in Eq. (1.1),

$\therefore \quad P(A \cap B) = P(A)\, P(B)$

Remark The additive law is used to find the probability of A or B, i.e., $P(A \cup B)$. The multiplicative law is used to find the probability of A and B, i.e., $P(A \cap B)$.

Corollary 1 If A, B and C are three events then

$$P(A\cap B\cap C)=P(A)\,P(B/A)\,P\left[C/(A\cap B)\right]$$

If A, B and C are independent events,

$$P(A\cap B\cap C)=P(A)\,P(B)\,P(C)$$

Corollary 2 If A and B are independent events then A and $\overline{B}$, $\overline{A}$ and B, $\overline{A}$ and $\overline{B}$ are also independent.

Corollary 3 The probability of occurrence of at least one of the events A, B, C is given by

$$P(A\cup B\cup C)=1-P(\overline{A}\cap\overline{B}\cap\overline{C})$$

If A, B, and C are independent events, their complements will also be independent.

$$P(A\cup B\cup C)=1-P(\overline{A})\,P(\overline{B})\,P(\overline{C})$$

Pairwise Independence and Mutual Independence The events A, B and C are mutually independent if the following conditions are satisfied simultaneously:

$$P(A\cap B)=P(A)\,P(B)$$
$$P(B\cap C)=P(B)\,P(C)$$
$$P(A\cap C)=P(A)\,P(C)$$

and $\quad P(A\cap B\cap C)=P(A)\,P(B)\,P(C)$

If the last condition is not satisfied, the events are said to be pairwise independent. Hence, mutually independent events are always pairwise independent but not vices versa.

Example 1

If A and B are two events such that $P(A)=\dfrac{2}{3}$, $P(\overline{A}\cap B)=\dfrac{1}{6}$ *and* $P(A\cap B)=\dfrac{1}{3}$, *find* $P(B)$, $P(A\cup B)$, $P(A/B)$, $P(B/A)$, $P(\overline{A}\cup B)$ *and* $P(\overline{B})$. *Also, examine whether the events A and B are (i) equally likely, (ii) exhaustive, (iii) mutually exclusive, and (iv) independent.*

Solution

$$\begin{aligned}P(B)&=P(\overline{A}\cap B)+P(A\cap B)\\&=\frac{1}{6}+\frac{1}{3}\\&=\frac{1}{2}\end{aligned}$$

$$\begin{aligned}P(A\cup B)&=P(A)+P(B)-P(A\cap B)\\&=\frac{2}{3}+\frac{1}{2}-\frac{1}{3}\\&=\frac{5}{6}\end{aligned}$$

$$\begin{aligned}P(A/B)&=\frac{P(A\cap B)}{P(B)}\\&=\frac{\left(\frac{1}{3}\right)}{\left(\frac{1}{2}\right)}\\&=\frac{2}{3}\end{aligned}$$

$$\begin{aligned}P(B/A)&=\frac{P(A\cap B)}{P(A)}\\&=\frac{\left(\frac{1}{3}\right)}{\left(\frac{2}{3}\right)}\\&=\frac{1}{2}\end{aligned}$$

$$\begin{aligned}P(\overline{A}\cup B)&=P(\overline{A})+P(B)-P(\overline{A}\cap B)\\&=\frac{1}{3}+\frac{1}{2}-\frac{1}{6}\\&=\frac{2}{3}\end{aligned}$$

$$\begin{aligned}P(\overline{A}\cap\overline{B})&=1-P(A\cup B)\\&=1-\frac{5}{6}\\&=\frac{1}{6}\end{aligned}$$

$$\begin{aligned}P(\overline{B})&=1-P(B)\\&=1-\frac{1}{2}\\&=\frac{1}{2}\end{aligned}$$

(i) Since $P(A)\neq P(B)$, A and B are not equally like events.

(ii) Since $P(A\cup B)\neq 1$, A and B are not exhaustive events.

(iii) Since $P(A \cap B) \neq 0$, A and B are not mutually exclusive events.

(iv) Since $P(A \cap B) = P(A)\, P(B)$, A and B are independent events.

Example 2

If A and B are two events such that $P(A) = 0.3$, $P(B) = 0.4$, $P(A \cap B) = 0.2$, *find (i)* $P(A \cup B)$, *(ii)* $P(\bar{A}/B)$, *and (iii)* $P(A/\bar{B})$.

Solution

(i)
$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= 0.3 + 0.4 - 0.2 \\ &= 0.5 \end{aligned}$$

(ii)
$$\begin{aligned} P(\bar{A}/B) &= \frac{P(\bar{A} \cap B)}{P(B)} \\ &= \frac{P(B) - P(A \cap B)}{P(B)} \\ &= \frac{0.4 - 0.2}{0.4} \\ &= 0.5 \end{aligned}$$

(iii)
$$\begin{aligned} P(A/\bar{B}) &= \frac{P(A \cap \bar{B})}{P(\bar{B})} \\ &= \frac{P(A) - P(A \cap B)}{1 - P(B)} \\ &= \frac{0.3 - 0.2}{1 - 0.4} \\ &= \frac{1}{6} \end{aligned}$$

Example 3

If A and B are two events with $P(A) = \frac{1}{3}, P(B) = \frac{1}{4}, P(A \cap B) = \frac{1}{12}$.

Find (i) $P(A/B)$, *(ii)* $P(B/A)$, *(iii)* $P(B/\bar{A})$, *and (iv)* $P(A \cap \bar{B})$.

Solution

(i)
$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{12}}{\frac{1}{4}} = \frac{1}{3}$$

(ii)
$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{1}{12}}{\frac{1}{3}} = \frac{1}{4}$$

(iii)
$$\begin{aligned} P(B/\bar{A}) &= \frac{P(B \cap \bar{A})}{P(\bar{A})} \\ &= \frac{P(B) - P(B \cap A)}{1 - P(A)} \\ &= \frac{\frac{1}{4} - \frac{1}{12}}{1 - \frac{1}{3}} \\ &= \frac{1}{4} \end{aligned}$$

(iv)
$$\begin{aligned} P(A \cap \bar{B}) &= P(A) - P(A \cap B) \\ &= \frac{1}{3} - \frac{1}{12} \\ &= \frac{1}{4} \end{aligned}$$

Example 4

Find the probability of drawing a queen and a king from a pack of cards in two consecutive draws, the cards drawn not being replaced.

Solution

Let A be the event that the card drawn is a queen.

$$P(A) = \frac{{}^{4}C_1}{{}^{52}C_1} = \frac{4}{52} = \frac{1}{13}$$

Let B be the event that the cards drawn are a king in the second draw given that the first card drawn is a queen.

$$P(B/A) = \frac{{}^{4}C_1}{{}^{51}C_1} = \frac{4}{51}$$

Probability that the cards drawn are a queen and a king

$$\begin{aligned} P(A \cap B) &= P(A)\, P(B/A) \\ &= \frac{4}{52} \times \frac{4}{51} \\ &= \frac{4}{663} \end{aligned}$$

Example 5

A bag contains 3 *red and* 4 *white balls. Two draws are made without replacement. What is the probability that both the balls are red?*

Solution

Let A be the event that the ball drawn is red in the first draw.

$$P(A) = \frac{3}{7}$$

Let B be the event that the ball drawn is red in the second draw given that the first ball drawn is red.

$$P(B/A) = \frac{2}{6}$$

Probability that both the balls are red

$$\begin{aligned} P(A \cap B) &= P(A)\,P(B/A) \\ &= \frac{3}{7} \times \frac{2}{6} \\ &= \frac{1}{7} \end{aligned}$$

Example 6

A bag contains 8 *red and* 5 *white balls. Two successive draws of* 3 *balls each are made such that (i) the balls are replaced before the second trial, and (ii) the balls are not replaced before the second trial. Find the probability that the first draw will give* 3 *white and the second,* 3 *red balls.*

Solution

Let A be the event that all 3 balls obtained at the first draw are white, and B be the event that all the 3 balls obtained at the second draw are red.

(i) When balls are replaced before the second trial,

$$P(A) = \frac{^{5}C_3}{^{13}C_3} = \frac{5}{143}$$

$$P(B) = \frac{^{8}C_3}{^{13}C_3} = \frac{28}{143}$$

Probability that the first draw will give 3 white and the second, 3 red balls

$$\begin{aligned} P(A \cap B) &= P(A)\,P(B) \\ &= \frac{5}{143} \times \frac{28}{143} \\ &= \frac{140}{20449} \end{aligned}$$

(ii) When the balls are not replaced before the second trial

$$P(B/A) = \frac{^{8}C_3}{10_{C_3}} = \frac{7}{15}$$

Probability that the first draw will give 3 white and the second, 3 red balls

$$\begin{aligned} P(A \cap B) &= P(A)\,P(B/A) \\ &= \frac{5}{143} \times \frac{7}{15} \\ &= \frac{7}{429} \end{aligned}$$

Example 7

From a bag containing 4 *white and* 6 *black balls, two balls are drawn at random. If the balls are drawn one after the other without replacements, find the probability that the first ball is white and the second ball is black.*

Solution

Let A be the event that the first ball drawn is white and B be the event that the second ball drawn is black given that the first ball drawn is white.

$$P(A) = \frac{4}{10}$$

$$P(B/A) = \frac{6}{9}$$

Probability that the first ball is white and the second ball is black.

$$\begin{aligned} P(A \cap B) &= P(A)\,P(B/A) \\ &= \frac{4}{10} \times \frac{6}{9} \\ &= \frac{4}{15} \end{aligned}$$

Example 8

Data on readership of a certain magazine show that the proportion of male readers under 35 is 0.40 and that over 35 is 0.20. If the proportion of readers under 35 is 0.70, find the probability of subscribers that are females over 35 years. Also, calculate the probability that a randomly selected male subscriber is under 35 years of age.

Solution

Let A be the event that the reader of the magazine is a male. Let B be the event that reader of the magazine is over 35 years of age.

$$P(A \cap \overline{B}) = 0.40, \qquad P(A \cap B) = 0.20, \qquad P(\overline{B}) = 0.7$$

$$\begin{aligned} P(B) &= 1 - P(\overline{B}) \\ &= 1 - 0.7 \\ &= 0.3 \end{aligned}$$

(i) Probability of subscribers that are females over 35 years

$$\begin{aligned} P(\overline{A} \cap B) &= P(B) - P(A \cap B) \\ &= 0.3 - 0.2 \\ &= 0.1 \end{aligned}$$

(ii) Probability that a randomly selected male subscriber is under 35 years of age

$$\begin{aligned} P(\overline{B}/A) &= \frac{P(A \cap \overline{B})}{P(A)} \\ &= \frac{P(A \cap \overline{B})}{P(A \cap B) + P(A \cap \overline{B})} \\ &= \frac{0.4}{0.2 + 0.4} \\ &= \frac{0.4}{0.6} \\ &= \frac{2}{3} \end{aligned}$$

Example 9

From a city population, the probability of selecting (a) a male or a smoker is $\frac{7}{10}$, *(b) a male smoker is* $\frac{2}{5}$, *and (c) a male, if a smoker is*

already selected, is $\frac{2}{3}$. *Find the probability of selecting (i) a nonsmoker, (ii) a male, and (iii) a smoker, if a male is first selected.*

Solution

Let A be the event that a male is selected. Let B be the event that a smoker is selected.

$$P(A \cup B) = \frac{7}{10}, \quad P(A \cap B) = \frac{2}{5}, \quad P(A/B) = \frac{2}{3}$$

(i) Probability of selecting a nonsmoker

$$\begin{aligned} P(\overline{B}) &= 1 - P(B) \\ &= 1 - \frac{P(A \cap B)}{P(A/B)} \\ &= 1 - \frac{\left(\frac{2}{5}\right)}{\left(\frac{2}{3}\right)} \\ &= \frac{2}{5} \end{aligned}$$

(ii)

$$\begin{aligned} P(B) &= 1 - P(\overline{B}) \\ &= 1 - \frac{2}{5} \\ &= \frac{3}{5} \end{aligned}$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad \text{...(1)}$$

Probability of selecting a male

$$\begin{aligned} P(A) &= P(A \cup B) + P(A \cap B) - P(B) \qquad \text{[Using Eq. (1)]} \\ &= \frac{7}{10} + \frac{2}{5} - \frac{3}{5} \\ &= \frac{1}{2} \end{aligned}$$

(iii) Probability of selecting a smoker if a male is first selected

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

$$= \frac{\left(\frac{2}{5}\right)}{\left(\frac{1}{2}\right)}$$

$$= \frac{4}{5}$$

Example 10

Sixty per cent of the employees of the XYZ corporation are college graduates. Of these, ten percent are in sales. Of the employee who did not graduate from college, eighty percent are in sales. What is the probability that

(i) an employee selected at random is in sales?

(ii) an employee selected at random is neither in sales nor a college graduate?

Solution

Let A be the event that an employee is a college graduate. Let B be the event that an employee is in sales.

$$P(A) = 0.6,\ \ P(B/A) = 0.10,\ \ P(B/\bar{A}) = 0.8$$

$$P(\bar{A}) = 1 - P(A) = 1 - 0.60 = 0.40$$

(i) Probability that an employee is in sales

$$\begin{aligned} P(B) &= P(A \cap B) + P(\bar{A} \cap B) \\ &= P(A)\,P(B/A) + P(\bar{A})\,P(B/\bar{A}) \\ &= (0.6 \times 0.1) + (0.40 \times 0.80) \\ &= 0.38 \end{aligned}$$

(ii) Probability that an employee is neither in sales nor a college graduate

$$\begin{aligned} P(\bar{A} \cap \bar{B}) &= 1 - P(A \cup B) \\ &= 1 - [P(A) + P(B) - P(A \cap B)] \\ &= 1 - [P(A) + P(B) - P(A)\,P(B/A)] \\ &= 1 - [0.60 + 0.38 - (0.60 \times 0.10)] \\ &= 0.08 \end{aligned}$$

Example 11

If A and B are two events such that $P(A) = \frac{3}{8}, P(B) = \frac{5}{8}$ *and* $P(A \cup B) = \frac{3}{4}$*, find* $P(A/B)$ *and* $P(B/A)$*. Show whether A and B are independent.*

Solution

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$\frac{3}{4} = \frac{3}{8} + \frac{5}{8} - P(A \cap B)$$

$$P(A \cap B) = \frac{1}{4}$$

$$\begin{aligned} P(A/B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{\left(\frac{1}{4}\right)}{\left(\frac{5}{8}\right)} \\ &= \frac{2}{5} \end{aligned}$$

$$\begin{aligned} P(B/A) &= \frac{P(A \cap B)}{P(A)} \\ &= \frac{\left(\frac{1}{4}\right)}{\left(\frac{3}{8}\right)} \\ &= \frac{2}{3} \end{aligned}$$

$$P(A)\,P(B) = \frac{3}{8} \times \frac{5}{8} = \frac{15}{64}$$

$$P(A \cap B) \neq P(A)\,P(B)$$

Hence, the events A and B are not independent.

Example 12

The probability that a student A solves a mathematics problem is $\frac{2}{5}$ *and the probability that a student B solves it is* $\frac{2}{3}$*. What is the probability* *that (i) the problem is not solved, (ii) the problem is solved, and (iii) both A and B, working independently of each other, solve the problem?*

Solution

Let A and B be events that students A and B solve the problem respectively.

$$P(A)=\frac{2}{5},\ P(B)=\frac{2}{3}$$

Events A and B are independent.

Probability that the student A does not solve the problem

$$\begin{aligned}P(\bar{A})&=1-P(A)\\&=1-\frac{2}{5}\\&=\frac{3}{5}\end{aligned}$$

Probability that the student B does not solve the problem

$$\begin{aligned}P(\bar{B})&=1-P(B)\\&=1-\frac{2}{3}\\&=\frac{1}{3}\end{aligned}$$

(i) Probability that the problem is not solved

$$\begin{aligned}P(\bar{A}\cap\bar{B})&=P(\bar{A})\,P(\bar{B})\\&=\frac{3}{5}\times\frac{1}{3}\\&=\frac{1}{5}\end{aligned}$$

(ii) Probability that the problem is solved

$$\begin{aligned}P(A\cup B)&=1-P(\bar{A}\cap\bar{B})\\&=1-\frac{1}{5}\\&=\frac{4}{5}\end{aligned}$$

(iii) Probability that both A and B solve the problem

$$\begin{aligned}P(A\cap B)&=P(A)\,P(B)\\&=\frac{2}{5}\times\frac{2}{3}\\&=\frac{4}{15}\end{aligned}$$

Example 13

The probability that the machine A will perform a usual function in 5 years' time is $\frac{1}{4}$, while the probability that the machine B will perform the function in 5 years' time is $\frac{1}{3}$. Find the probability that both machines will perform the usual function.

Solution

Let A and B be the events that machines A and B will perform the usual function respectively.

$$P(A)=\frac{1}{4}$$

$$P(B)=\frac{1}{3}$$

Events A and B are independent.

Probability that both machines will perform the usual function

$$\begin{aligned}P(A\cap B)&=P(A)\,P(B)\\&=\frac{1}{4}\times\frac{1}{3}\\&=\frac{1}{12}\end{aligned}$$

Example 14

A person A is known to hit a target in 3 out of 4 shots, whereas another person B is known to hit the same target in 2 out of 3 shots. Find the probability of the target being hit at all when they both try.

[Summer 2015]

Solution

Let A and B be the events that the persons A and B hit the target respectively.

$$P(A)=\frac{3}{4}$$

$$P(B)=\frac{2}{3}$$

Events A and B are independent.

Probability that the person A will not hit the target $=P(\bar{A})=1-P(A)=1-\frac{3}{4}=\frac{1}{4}$

Probability that the person B will not hit the target $= P(\bar{B}) = 1 - P(B) = 1 - \frac{2}{3} = \frac{1}{3}$

Probability that the target is not hit at all

$$P(\bar{A} \cap \bar{B}) = P(\bar{A})\,P(\bar{B})$$
$$= \frac{1}{4} \times \frac{1}{3}$$
$$= \frac{1}{12}$$

Probability that the target is hit at all when they both try

$$P(A \cup B) = 1 - P(\bar{A} \cap \bar{B})$$
$$= 1 - \frac{1}{12}$$
$$= \frac{11}{12}$$

Aliter

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$= P(A) + P(B) - P(A)\,P(B) \quad [\because \ A \text{ and } B \text{ independent}]$$
$$= \frac{3}{4} + \frac{2}{3} - \frac{3}{4} \times \frac{2}{3}$$
$$= \frac{11}{12}$$

Example 15

The odds against A speaking the truth are 4 : 6 *while the odds in favour of B speaking the truth are* 7 : 3. *What is the probability that A and B contradict each other in stating the same fact?*

Solution

Let A and B be events that A and B speak the truth respectively.

$$P(A) = \frac{6}{10}$$
$$P(B) = \frac{7}{10}$$

Events A and B are independent.

Probability that A speaks a lie $= P(\bar{A}) = 1 - P(A) = 1 - \frac{6}{10} = \frac{4}{10}$

Probability that B speaks a lie $= P(\bar{B}) = 1 - P(B) = 1 - \frac{7}{10} = \frac{3}{10}$

Probability that A and B contradict each other

$$P[(A \cap \bar{B}) \cup (\bar{A} \cap B)] = P(A \cap \bar{B}) + P(\bar{A} \cap B) \quad \left[\because \ (A \cap \bar{B}) \text{ and } (\bar{A} \cap B) \text{ are mutually exclusive events}\right]$$
$$= P(A)\,P(\bar{B}) + P(\bar{A})\,P(B)$$
$$= \frac{6}{10} \times \frac{3}{10} + \frac{4}{10} \times \frac{7}{10}$$
$$= \frac{23}{50}$$

Example 16

An urn contains 10 *red,* 5 *white and* 5 *blue balls. Two balls are drawn at random. Find the probability that they are not of the same colour.*

Solution

Let A, B, and C be the events that two balls drawn at random be of the same colour, i.e., red, white, and blue respectively.

$$P(A) = \frac{{}^{10}C_2}{{}^{20}C_2} = \frac{9}{38}$$
$$P(B) = \frac{{}^{5}C_2}{{}^{20}C_2} = \frac{1}{19}$$
$$P(C) = \frac{{}^{5}C_2}{{}^{20}C_2} = \frac{1}{19}$$

Events A, B, and C are independent.

Probability that both balls drawn are of same colour

$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$
$$= \frac{9}{38} + \frac{1}{19} + \frac{1}{19}$$
$$= \frac{13}{38}$$

Probability that both balls drawn are not of the same colour

$$P(\bar{A} \cap \bar{B} \cap \bar{C}) = 1 - P(A \cup B \cup C)$$
$$= 1 - \frac{13}{38}$$
$$= \frac{25}{38}$$

Example 17

A problem in statistics is given to three students A, B and C, whose chances of solving it independently are $\frac{1}{2}, \frac{1}{3}$, and $\frac{1}{4}$ respectively. Find the probability that

(i) the problem is solved
(ii) at least two of them are able to solve the problem
(iii) exactly two of them are able to solve the problem
(iv) exactly one of them is able to solve the problem

Solution

Let A, B, and C be the events that students A, B, and C solve the problem respectively.

$$P(A)=\frac{1}{2},\ P(B)=\frac{1}{3},\ P(C)=\frac{1}{4}$$

Events A, B, and C are independent.

(i) Probability that the problem is solved or at least one of them is able to solve the problem is same.

$$\begin{aligned}P(A\cup B\cup C) &= P(A)+P(B)+P(C)-P(A\cap B)-P(A\cap C)-P(B\cap C)\\ &\qquad +P(A\cap B\cap C)\\ &= P(A)+P(B)+P(C)-P(A)\,P(B)-P(A)\,P(C)-P(B)\,P(C)\\ &\qquad +P(A)\,P(B)\,P(C)\\ &= \frac{1}{2}+\frac{1}{3}+\frac{1}{4}-\left(\frac{1}{2}\times\frac{1}{3}\right)-\left(\frac{1}{2}\times\frac{1}{4}\right)-\left(\frac{1}{3}\times\frac{1}{4}\right)+\left(\frac{1}{2}\times\frac{1}{3}\times\frac{1}{4}\right)\\ &= \frac{3}{4}\end{aligned}$$

(ii) Probability that at least two of them are able to solve the problem

$$\begin{aligned}P[(A\cap B)\cup(B\cap C)\cup(A\cap C)] &= P(A\cap B)+P(B\cap C)+P(A\cap C)-2P(A\cap B\cap C)\\ &= P(A)\,P(B)+P(B)\,P(C)+P(A)\,P(C)\\ &\qquad -2P(A)\,P(B)\,P(C)\\ &= \left(\frac{1}{2}\times\frac{1}{3}\right)+\left(\frac{1}{3}\times\frac{1}{4}\right)+\left(\frac{1}{2}\times\frac{1}{4}\right)-2\left(\frac{1}{2}\times\frac{1}{3}\times\frac{1}{4}\right)\\ &= \frac{7}{24}\end{aligned}$$

(iii) Probability that exactly two of them are able to solve the problem

$$\begin{aligned}&P\left[(A\cap B\cap\bar{C})\cup(A\cap\bar{B}\cap C\cup(\bar{A}\cap B\cap C\right]\\ &= P(A\cap B)+P(B\cap C)+P(A\cap C)-3P(A\cap B\cap C)\\ &= P(A)\,P(B)+P(B)\,P(C)+P(A)\,P(C)-3P(A)\,P(B)\,P(C)\\ &= \left(\frac{1}{2}\times\frac{1}{3}\right)+\left(\frac{1}{3}\times\frac{1}{4}\right)+\left(\frac{1}{2}\times\frac{1}{4}\right)-3\left(\frac{1}{2}\times\frac{1}{3}\times\frac{1}{4}\right)\\ &= \frac{1}{4}\end{aligned}$$

(iv) Probability that exactly one of them is able to solve the problem

$$\begin{aligned}&P\left[A\cap\bar{B}\cap\bar{C})\cup(\bar{A}\cap B\cap\bar{C})\cup(\bar{A}\cap\bar{B}\cap C)\right]\\ &= P(A)+P(B)+P(C)-2P(A\cap B)-2P(B\cap C)-2P(A\cap C)+3P(A\cap B\cap C)\\ &= \frac{1}{2}+\frac{1}{3}+\frac{1}{4}-2\left(\frac{1}{2}\times\frac{1}{3}\right)-2\left(\frac{1}{3}\times\frac{1}{4}\right)-2\left(\frac{1}{2}\times\frac{1}{4}\right)+3\left(\frac{1}{2}\times\frac{1}{3}\times\frac{1}{4}\right)\\ &= \frac{11}{24}\end{aligned}$$

Example 18

A husband and wife appeared in an interview for two vacancies in an office. The probability of the husband's selection is $\frac{1}{7}$ and that of the wife's selection is $\frac{1}{5}$. Find the probability that (i) both of them are selected, (ii) only one of them is selected, (iii) none of them is selected, and (iv) at least one of them is selected.

Solution

Let A and B be the events that the husband and wife are selected respectively.

$$P(A)=\frac{1}{7},\qquad P(B)=\frac{1}{5}.$$

Events A and B are independent.

(i) Probability that both of them are selected

$$\begin{aligned}P(A\cap B) &= P(A)\,P(B)\\ &= \frac{1}{7}\times\frac{1}{5}\\ &= \frac{1}{35}\end{aligned}$$

(ii) Probability that at least one of them is selected

$$\begin{aligned} P(A\cup B) &= P(A)+P(B)-P(A\cap B)\\ &= \frac{1}{7}+\frac{1}{5}-\frac{1}{35}\\ &= \frac{11}{35} \end{aligned}$$

(iii) Probability that none of them is selected

$$\begin{aligned} P(\bar{A}\cap\bar{B}) &= 1-P(A\cup B)\\ &= 1-\frac{11}{35}\\ &= \frac{24}{35} \end{aligned}$$

(iv) Probability that only one of them is selected

$$\begin{aligned} P\left[A\cap\bar{B})\cup(\bar{A}\cap B)\right] &= P(A\cup B)-P(A\cap B)\\ &= \frac{11}{35}-\frac{1}{35}\\ &= \frac{10}{35}\\ &= \frac{2}{7} \end{aligned}$$

Example 19

There are two bags. The first contains 2 red and 1 white ball, whereas the second bag has only 1 red and 2 white balls. One ball is taken out at random from the first bag and put in the second. Then a ball is chosen at random from the second bag. What is the probability that this last ball is red?

Solution

There are two mutually exclusive cases.

Case I: A red ball is transferred from the first bag to the second bag and a red ball is drawn from it.

Case II: A white ball is transferred from the first bag to the second bag and then a red ball is drawn from it.

Let A be the event of transferring a red ball from the first bag, and B be the event of transferring a white ball from the first bag.

$$P(A)=\frac{2}{3}$$

$$P(B)=\frac{1}{3}$$

Let E be the event of drawing a red ball from the second bag.

$$P(E/A)=\frac{2}{4}$$

$$P(E/B)=\frac{1}{4}$$

$$\begin{aligned} P(\text{Case I}) &= P(A\cap E)\\ &= P(A)\,P(E/A)\\ &= \frac{2}{3}\times\frac{2}{4}\\ &= \frac{1}{3} \end{aligned}$$

$$\begin{aligned} P(\text{Case II}) &= P(B\cap E)\\ &= P(B)\,P(E/B)\\ &= \frac{1}{3}\times\frac{1}{4}\\ &= \frac{1}{12} \end{aligned}$$

$$\begin{aligned} P\left[(A\cap E)\cup(B\cap E)\right] &= P(A\cap E)+P(B\cap E)\\ &= \frac{1}{3}+\frac{1}{12}\\ &= \frac{5}{12} \end{aligned}$$

Example 20

An urn contains four tickets marked with numbers 112, 121, 211, and 222, and one ticket is drawn. Let A_i $(i = 1, 2, 3)$ be the event that the i^{th} digit of the ticket drawn is 1. Show that the events A_1, A_2, A_3 are pairwise independent but not mutually independent.

Solution

$$A_1=\{112,121\},\ A_2=\{112,211\},\ A_3=\{121,211\}$$

$$A_1\cap A_2=\{112\},\ A_1\cap A_3=\{121\},\ A_2\cap A_3=\{211\}$$

$$P(A_1)=\frac{2}{4}=\frac{1}{2}=P(A_2)=P(A_3)$$

$$P(A_1\cap A_2)=\frac{1}{4}=P(A_1\cap A_3)=P(A_2\cap A_3)$$

$$P(A_1 \cap A_2) = P(A_1)\,P(A_2) = \frac{1}{4}$$

$$P(A_2 \cap A_3) = P(A_2)\,P(A_3) = \frac{1}{4}$$

$$P(A_1 \cap A_3) = P(A_1)\,P(A_3) = \frac{1}{4}$$

Hence, events A_1, A_2, and A_3 are pairwise independent.

$$P(A_1 \cap A_2 \cap A_3) = P(\phi) = 0$$

$$P(A_1 \cap A_2 \cap A_3) \neq P(A_1)\,P(A_2)\,P(A_3)$$

Hence, events A_1, A_2, and A_3 are not mutually independent.

EXERCISE 1.3

1. Find the probability of drawing 2 red balls in succession from a bag containing 4 red and 5 black balls when the ball that is drawn first is (i) not replaced, and (ii) replaced.

[**Ans.:** (i) $\frac{1}{6}$ (ii) $\frac{16}{81}$]

2. Two aeroplanes bomb a target in succession. The probability of each correctly scoring a hit is 0.3 and 0.2 respectively. The second will bomb only if the first misses the target. Find the probability that (i) the target is hit, and (ii) both fail to score hits.

[**Ans.:** (i) 0.44 (ii) 0.56]

3. Box *A* contains 5 red and 3 white marbles and Box *B* contains 2 red and 6 white marbles. If a marble is drawn from each box, what is the probability that they are both of the same colour?

[**Ans.:** 0.109]

4. Two marbles are drawn in succession from a box containing 10 red, 30 white, 20 blue, and 15 orange marbles, with replacement being made after each draw. Find the probability that (i) both are white, and (ii) the first is red and the second is white.

[**Ans.:** (i) $\frac{4}{25}$ (ii) $\frac{4}{75}$]

5. *A*, *B*, *C* are aiming to shoot a balloon. *A* will succeed 4 times out of 5 attempts. The chance of *B* to shoot the balloon is 3 out of 4, and that of *C* is 2 out of 3. If the three aim the balloon simultaneously, find the probability that at least two of them hit the balloon.

[**Ans.:** $\frac{5}{6}$]

6. There are 12 cards numbered 1 to 12 in a box. If two cards are selected, what is the probability that the sum is odd (i) with replacement, and (ii) without replacement?

[**Ans.:** (i) $\frac{1}{2}$ (ii) $\frac{6}{11}$]

7. Two cards are drawn from a well-shuffled pack of 52 cards. Find the probability that they are both aces if the first card is (i) replaced, and (ii) not replaced.

[**Ans.:** (i) $\frac{1}{169}$ (ii) $\frac{1}{221}$]

8. *A* can hit a target 2 times in 5 shots; B, 3 times in 4 shots; and C, 2 times in 3 shots. They fire a volley. What is the probability that at least 2 shots hit the target?

[**Ans.:** $\frac{2}{3}$]

9. There are two bags. The first bag contains 5 red and 7 white balls and the second bag contains 3 red and 12 white balls. One ball is taken out at random from the first bag and is put in the second bag. Now, a ball is drawn from the second bag. What is the probability that this last ball is red?

[**Ans.:** $\frac{41}{192}$]

10. In a shooting competition, the probability of *A* hitting the target is $\frac{1}{2}$; of *B*, is $\frac{2}{3}$; and of C, is $\frac{3}{4}$. If all of them fire at the target, find the probability that (i) none of them hits the target, and (ii) at least one of them hits the target.

[**Ans.:** (i) $\frac{1}{24}$ (ii) $\frac{23}{24}$]

11. The odds against a student *X* solving a statistics problem are 12 to 10 and the odds in favour of a student *Y* solving the problem are 6 to 9.

What is the probability that the problem will be solved when both try independently of each other?

$\left[\text{Ans.: } \frac{37}{55}\right]$

12. A bag contains 6 white and 9 black balls. Four balls are drawn at random twice. Find the probability that the first draw will give 4 white balls and the second draw will give 4 black balls if (i) the balls are replaced, and (ii) the balls are not replaced before the second draw.

$\left[\text{Ans.: (i) } \frac{6}{5915} \text{ (ii) } \frac{3}{715}\right]$

13. An urn contains 10 white and 3 black balls. Another urn contains 3 white and 5 black balls. Two balls are transferred from the first urn to the second urn and then one ball is drawn from the latter. What is the probability that the ball drawn is white?

$\left[\text{Ans.: } \frac{5}{26}\right]$

14. A man wants to marry a girl having the following qualities: fair complexion—the probability of getting such a girl is $\frac{1}{20}$, handsome dowry—the probability is $\frac{1}{50}$, westernized manners and etiquettes—the probability of this is $\frac{1}{100}$. Find the probability of his getting married to such a girl when the possessions of these three attributes are independent.

$\left[\text{Ans.: } \frac{1}{100000}\right]$

15. A small town has one fire engine and one ambulance available for emergencies. The probability that the fire engine is available when needed is 0.98 and the probability that the ambulance is available when called is 0.92. In the event of an injury resulting from a burning building, find the probability that both the fire engine and ambulance will be available.

[Ans.: 0.9016]

16. In a certain community, 36% of the families own a dog and 22% of the families that own a dog also own a cat. In addition, 30% of the families own a cat. What is the probability that (i) a randomly selected family owns both a dog and a cat, and (ii) a randomly selected family owns a dog given that it owns a cat?

[Ans.: (i) 0.0792 (ii) 0.264]

1.7 BAYES' THEOREM

Let $A_1, A_2, \ldots, A_n$ be n mutually exclusive and exhaustive events with $P(A_i) \neq 0$ for $i = 1, 2, \ldots, n$ in a sample space S. Let B be an event that can occur in combination with any one of the events $A_1, A_2, \ldots, A_n$ with $P(B) \neq 0$. The probability of the event A_i when the event B has actually occurred is given by

$$P(A_i/B) = \frac{P(A_i)\,P(B/A_i)}{\sum_{i=1}^{n} P(A_i)\,P(B/A_i)}$$

Proof Since $A_1, A_2, \ldots, A_n$ are n mutually exclusive and exhaustive events of the sample space S,

$$S = A_1 \cup A_2 \cup \ldots \cup A_n$$

Since B is another event that can occur in combination with any of the mutually exclusive and exhaustive events $A_1, A_2, \ldots, A_n$,

$$B = (A_1 \cap B) \cup (A_2 \cap B) \cup \cdots \cup (A_n \cap B)$$

Taking probability of both the sides,

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + \cdots + P(A_n \cap B)$$

The events $(A_1 \cap B), (A_2 \cap B)$, etc., are mutually exclusive.

$$P(B) = \sum_{i=1}^{n} P(A_i \cap B) = \sum_{i=1}^{n} P(A_i)\,P(B/A_i)$$

The conditional probability of an event A given that B has already occurred is given by

$$\begin{aligned} P(A_i/B) &= \frac{P(A_i \cap B)}{P(B)} \\ &= \frac{P(A_i)\,P(B/A_i)}{P(B)} \\ &= \frac{P(A_i)\,P(B/A_i)}{\sum_{i=1}^{n} P(A_i)\,P(B/A_i)} \end{aligned}$$

Example 1

A company has two plants to manufacture hydraulic machines. Plant I manufactures 70% of the hydraulic machines, and Plant II manufactures 30%. At Plant I, 80% of hydraulic machines are rated standard quality; and at Plant II, 90% of hydraulic machines are rated standard quality. A machine is picked up at random and is found to be of standard quality. What is the chance that it has come from Plant I? **[Summer 2015]**

Solution

Let A_1 and A_2 be the events that the hydraulic machines are manufactured in Plant I and Plant II respectively. Let B be the event that the machine picked up is found to be of standard quality.

$$P(A_1)=\frac{70}{100}=0.7$$

$$P(A_2)=\frac{30}{100}=0.3$$

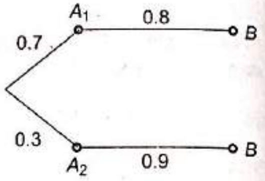


Fig. 1.3

Probability that the machine is of standard quality given that it is manufactured in Plant I

$$P(B/A_1)=\frac{80}{100}=0.8$$

Probability that the machine is of standard quality given that it is manufactured in Plant II

$$P(B/A_2)=\frac{90}{100}=0.9$$

Probability that a machine is manufactured in Plant I given that it is of standard quality

$$P(A_1/B)=\frac{P(A_1)\,P(B/A_1)}{P(A_1)\,P(B/A_1)+P(A_2)\,P(B/A_2)}$$

$$=\frac{0.7\times0.8}{0.7\times0.8+0.3\times0.9}$$

$$=0.6747$$

Example 2

A bag A contains 2 white and 3 red balls, and a bag B contains 4 white and 5 red balls. One ball is drawn at random from one of the bags and it is found to be red. Find the probability that the red ball is drawn from the bag B.

Solution

Let A_1 and A_2 be the events that the ball is drawn from bags A and B respectively. Let B be the event that the ball drawn is red.

$$P(A_1)=\frac{1}{2}$$

$$P(A_2)=\frac{1}{2}$$

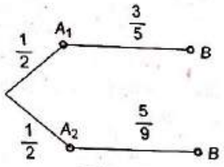


Fig. 1.4

Probability that the ball drawn is red given that it is drawn from the bag A

$$P(B/A_1)=\frac{3}{5}$$

Probability that the ball drawn is red given that it is drawn from the bag B

$$P(B/A_2)=\frac{5}{9}$$

Probability that the ball is drawn from the bag B given that it is red

$$P(A_2/B)=\frac{P(A_2)\,P(B/A_2)}{P(A_1)\,P(B/A_1)+P(A_2)\,P(B/A_2)}$$

$$=\frac{\frac{1}{2}\times\frac{5}{9}}{\left(\frac{1}{2}\times\frac{3}{5}\right)+\left(\frac{1}{2}\times\frac{5}{9}\right)}$$

$$=\frac{25}{52}$$

Example 3

The chances that Doctor A will diagnose a disease X correctly is 60%. The chances that a patient will die by his treatment after correct diagnosis is 40% and the chance of death by wrong diagnosis is 70%. A patient of Doctor A, who had the disease X, died. What is the chance that his disease was diagnosed correctly?

Solution

Let A_1 be the event that the disease X is diagnosed correctly by Doctor A. Let A_2 be the event that the disease X is not diagnosed correctly by Doctor A. Let B be the event that a patient of Doctor A who has the disease X, dies.

$$P(A_1)=\frac{60}{100}=0.6$$

$$P(A_2)=P(\bar{A}_1)=1-P(A_1)=0.4$$

Probability that the patient of Doctor *A* who has the disease *X* dies given that the disease *X* is diagnosed correctly

$$P(B/A_1)=\frac{40}{100}=0.4$$

Probability that the patient of Doctor *A* who has the disease *X* dies given that the disease *X* is not diagnosed correctly

$$P(B/A_2)=\frac{70}{100}=0.7$$

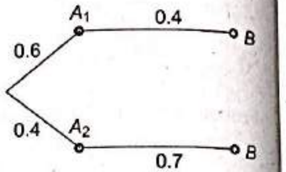


Fig. 1.5

Probability that the disease *X* is diagnosed correctly given that a patient of Doctor *A* who has the disease *X* dies

$$P(A_1/B)=\frac{P(A_1)\,P(B/A_1)}{P(A_1)\,P(B/A_1)+P(A_2)\,P(B/A_2)}$$

$$=\frac{0.6\times0.4}{(0.6\times0.4)+(0.4\times0.7)}$$

$$=\frac{6}{13}$$

Example 4

In a bolt factory, machines A, B, C manufacture 25%, 35%, and 40% of the total output and out of the total manufacturing, 5%, 4%, and 2% are defective bolts. A bolt is drawn at random from the product and is found to be defective. Find the probabilities that it is manufactured from (i) Machine A, (ii) Machine B, and (iii) Machine C.

Solution

Let A_1, A_2 and A_3 be the events that bolts are manufactured by machines *A*, *B*, and *C* respectively. Let *B* be the event that the bolt drawn is defective.

$$P(A_1)=\frac{25}{100}=0.25$$

$$P(A_2)=\frac{35}{100}=0.35$$

$$P(A_3)=\frac{40}{100}=0.4$$

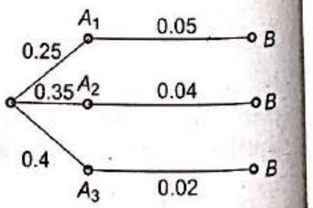


Fig. 1.6

Probability that the bolt drawn is defective given that it is manufactured from Machine *A*

$$P(B/A_1)=\frac{5}{100}=0.05$$

Probability that the bolt drawn is defective given that it is manufactured from Machine *B*

$$P(B/A_2)=\frac{4}{100}=0.04$$

Probability that the bolt drawn is defective given that it is manufactured from Machine *C*

$$P(B/A_3)=\frac{2}{100}=0.02$$

(i) Probability that a bolt is manufactured from Machine *A* given that it is defective

$$P(A_1/B)=\frac{P(A_1)\,P(B/A_1)}{P(A_1)\,P(B/A_1)+P(A_2)\,P(B/A_2)+P(A_3)\,P(B/A_3)}$$

$$=\frac{0.25\times0.05}{(0.25\times0.05)+(0.35\times0.04)+(0.4\times0.02)}$$

$$=0.3623$$

(ii) Probability that a bolt is manufactured from Machine *B* given that it is defective

$$P(A_2/B)=\frac{P(A_2)\,P(B/A_2)}{P(A_1)\,P(B/A_1)+P(A_2)\,P(B/A_2)+P(A_3)\,P(B/A_3)}$$

$$=\frac{0.35\times0.04}{(0.25\times0.05)+(0.35\times0.04)+(0.4\times0.02)}$$

$$=0.4058$$

(iii) Probability that a bolt is manufactured from Machine C given that it is defective

$$P(A_3/B)=\frac{P(A_3)\,P(B/A_3)}{P(A_1)\,P(B/A_1)+P(A_2)\,P(B/A_2)+P(A_3)\,P(B/A_3)}$$

$$=\frac{0.4\times0.02}{(0.25\times0.05)+(0.35\times0.04)+(0.4\times0.02)}$$

$$=0.2319$$

Example 5

A businessman goes to hotels X, Y, Z for 20%, 50%, 30% of the time respectively. It is known that 5%, 4%, 8% of the rooms in X, Y, Z hotels have faulty plumbings. What is the probability that the businessman's room having faulty plumbing is assigned to Hotel Z?

Solution

Let A_1, A_2 and A_3 be the events that the businessman goes to hotels X, Y, Z respectively. Let B be the event that the rooms have faulty plumbings.

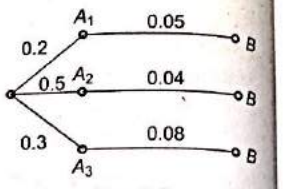


Fig. 1.7

$$P(A_1) = \frac{20}{100} = 0.2$$

$$P(A_2) = \frac{50}{100} = 0.5$$

$$P(A_3) = \frac{30}{100} = 0.3$$

Probability that rooms have faulty plumbings given that rooms belong to Hotel X

$$P(B/A_1) = \frac{5}{100} = 0.05$$

Probability that rooms have faulty plumbing given that rooms belong to Hotel Y

$$P(B/A_2) = \frac{4}{100} = 0.04$$

Probability that rooms have faulty plumbings given that rooms belong to Hotel Z

$$P(B/A_3) = \frac{8}{100} = 0.08$$

Probability that the businessman's room belongs to Hotel Z given that the room has faulty plumbing

$$P(A_3/B) = \frac{P(A_3)\,P(B/A_3)}{P(A_1)\,P(B/A_1) + P(A_2)\,P(B/A_2) + P(A_3)\,P(B/A_3)}$$

$$= \frac{0.3 \times 0.08}{(0.2 \times 0.05) + (0.5 \times 0.04) + (0.3 \times 0.08)}$$

$$= \frac{4}{9}$$

Example 6

Of three persons the chances that a politician, a businessman, or an academician would be appointed the Vice Chancellor (VC) of a university are 0.5, 0.3, 0.2 *respectively. Probabilities that research is promoted by these persons if they are appointed as VC are* 0.3, 0.7, 0.8 *respectively.*

(i) Determine the probability that research is promoted.

(ii) If research is promoted, what is the probability that the VC is an academician?

Solution

Let A_1, A_2 and A_3 be the events that a politician, a businessman or an academician will be appointed as the VC respectively. Let B be the event that research is promoted by these persons if they are appointed as VC.

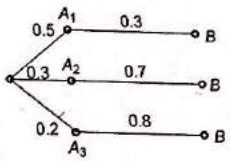


Fig. 1.8

$$P(A_1) = 0.5$$

$$P(A_2) = 0.3$$

$$P(A_3) = 0.2$$

Probability that research is promoted given that a politician is appointed as VC

$$P(B/A_1) = 0.3$$

Probability that research is promoted given that a businessman is promoted as VC

$$P(B/A_2) = 0.7$$

Probability that research is promoted given that an academician is appointed as VC

$$P(B/A_3) = 0.8$$

(i) Probability that research is promoted

$$P(B) = P(A_1)\,P(B/A_1) + P(A_2)\,P(B/A_2) + P(A_3)\,P(B/A_3)$$

$$= (0.5 \times 0.3) + (0.3 \times 0.7) + (0.2 \times 0.8)$$

$$= 0.52$$

(ii) Probability that the VC is an academician given that research is promoted by him

$$P(A_3/B) = \frac{P(A_3)\,P(B/A_3)}{P(A_1)\,P(B/A_1) + P(A_2)\,P(B/A_2) + P(A_3)\,P(B/A_3)}$$

$$= \frac{0.2 \times 0.8}{0.52}$$

$$= \frac{4}{13}$$

Example 7

The contents of urns I, II, and III are as follows:

1 white, 2 red, and 3 black balls,

2 white, 3 red, and 1 black ball, and

3 white, 1 red, and 2 black balls.

One urn is chosen at random and two balls are drawn. They happen to be white and red. Find the probability that they came from (i) Urn I, (ii) Urn II, and (iii) Urn III.

Solution

Let A_1, A_2, and A_3 be the events that urns I, II and III are chosen respectively. Let B be the event that 2 balls drawn are white and red.

$$P(A_1)=\frac{1}{3}$$

$$P(A_2)=\frac{1}{3}$$

$$P(A_3)=\frac{1}{3}$$

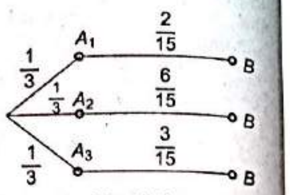


Fig. 1.9

Probability that 2 balls drawn are white and red given that they are chosen from the urn I

$$P(B/A_1)=\frac{{}^1C_1\times{}^2C_1}{{}^6C_2}=\frac{1\times2}{15}=\frac{2}{15}$$

Probability that 2 balls drawn are white and red given that they are chosen from the urn II

$$P(B/A_2)=\frac{{}^2C_1\times{}^3C_1}{{}^6C_2}=\frac{2\times3}{15}=\frac{6}{15}$$

Probability that 2 balls drawn are white and red given that they are chosen from the urn III

$$P(B/A_3)=\frac{{}^3C_1\times{}^1C_1}{{}^6C_2}=\frac{3\times1}{15}=\frac{3}{15}$$

(i) Probability that 2 balls came from the urn I given that they are white and red

$$P(A_1/B)=\frac{P(A_1)\,P(B/A_1)}{P(A_1)\,P(B/A_1)+P(A_2)\,P(B/A_2)+P(A_3)\,P(B/A_3)}$$

$$=\frac{\frac{1}{3}\times\frac{2}{15}}{\left(\frac{1}{3}\times\frac{2}{15}\right)+\left(\frac{1}{3}\times\frac{6}{15}\right)+\left(\frac{1}{3}\times\frac{3}{15}\right)}$$

$$=\frac{2}{11}$$

(ii) Probability that 2 balls came from the urn II given that they are white and red

$$P(A_2/B)=\frac{P(A_2)\,P(B/A_2)}{P(A_1)\,P(B/A_1)+P(A_2)\,P(B/A_2)+P(A_3)\,P(B/A_3)}$$

$$=\frac{\frac{1}{3}\times\frac{6}{15}}{\left(\frac{1}{3}\times\frac{2}{15}\right)+\left(\frac{1}{3}\times\frac{6}{15}\right)+\left(\frac{1}{3}\times\frac{3}{15}\right)}$$

$$=\frac{6}{11}$$

(iii) Probability that 2 balls came from the urn III given that they are white and red

$$P(A_3/B)=\frac{P(A_3)\,P(B/A_3)}{P(A_1)\,P(B/A_1)+P(A_2)\,P(B/A_2)+P(A_3)\,P(B/A_3)}$$

$$=\frac{\frac{1}{3}\times\frac{3}{15}}{\left(\frac{1}{3}\times\frac{2}{15}\right)+\left(\frac{1}{3}\times\frac{6}{15}\right)+\left(\frac{1}{3}\times\frac{3}{15}\right)}$$

$$=\frac{3}{11}$$

EXERCISE 1.4

1. There are 4 boys and 2 girls in Room A and 5 boys and 3 girls in Room B. A girl from one of the two rooms laughed loudly. What is the probability the girl who laughed was from Room B?

   $\left[\textbf{Ans.:}\ \frac{9}{17}\right]$

2. The probability of X, Y, and Z becoming managers are $\frac{4}{9}, \frac{2}{9}$, and $\frac{1}{3}$ respectively. The probabilities that the bonus scheme will be introduced if X, Y, and Z become managers are $\frac{3}{10}, \frac{1}{2}$, and $\frac{4}{5}$ respectively. (i) What is the probability that the bonus scheme will be introduced? (ii) If the bonus scheme has been introduced, what is the probability that the manager appointed was X?

   $\left[\textbf{Ans.:}\ \text{(i)}\ \frac{23}{45}\ \ \text{(ii)}\ \frac{6}{23}\right]$

3. A factory has two machines, A and B. Past records show that the machine A produces 30% of the total output and the machine B, the remaining 70%. Machine A produces 5% defective articles and Machine B produces 1% defective items. An item is drawn at random and found to be defective. What is the probability that it was produced (i) by the machine A, and (ii) by the Machine B?

   [**Ans.:** (i) 0.682 (ii) 0.318]

4. A company has two plants to manufacture scooters. Plant I manufactures 80% of the scooters, and Plant II manufactures 20%. At Plant I, 85 out of 100 scooters are rated standard quality or better. At Plant II, only 65 out of 100 scooters are rated standard quality or better. What is the probability that a scooter selected at random came from (i) Plant I, and (ii) Plant II if it is known that the scooter is of standard quality?

[**Ans.:** (i) 0.84 (ii) 0.16]

5. A new pregnancy test was given to 100 pregnant women and 100 non-pregnant women. The test indicated pregnancy in 92 of the 100 pregnant women and in 12 of the 100 non-pregnant women. If a randomly selected woman takes this test and the test indicates she is pregnant, what is the probability she was not pregnant?

$\left[\textbf{Ans.: } \frac{3}{26}\right]$

6. An insurance company insured 2000 scooter drivers, 4000 car drivers, and 6000 truck drivers. The probability of an accident is 0.01, 0.03, and 0.15 in the respective category. One of the insured drivers meets with an accident. What is the probability that he is a scooter driver?

$\left[\textbf{Ans.: } \frac{1}{52}\right]$

7. Consider a population of consumers consisting of two types. The upper-income class of consumers comprise 35% of the population and each member has a probability of 0.8 of purchasing Brand *A* of a product. Each member of the rest of the population has a probability of 0.3 of purchasing Brand *A* of the product. A consumer, chosen at random, is found to be the buyer of Brand *A*. What is the probability that the buyer belongs to the middle-income and lower-income classes of consumers?

$\left[\textbf{Ans.: } \frac{39}{95}\right]$

8. There are two boxes of identical appearance, each containing 4 spark plugs. It is known that the box I contains only one defective spark plug, while all the four spark plugs of the box II are non-defective. A spark plug drawn at random from a box, selected at random, is found to be non-defective. What is the probability that it came from the box I?

$\left[\textbf{Ans.: } \frac{3}{7}\right]$

9. Vijay has 5 one-rupee coins and one of them is known to have two heads. He takes out a coin at random and tosses it 5 times—it always falls head upward. What is the probability that it is a coin with two heads?

$\left[\textbf{Ans.: } \frac{8}{9}\right]$

10. Stores *A*, *B*, and *C* have 50, 75, and 100 employees and, respectively 50, 60, 70 per cent of these are women. Resignations are equally likely among all employees, regardless of sex. One employee resigns and this is a woman. What is the probability that she works in Store *C*?

[**Ans.:** 0.5]

CHAPTER 2

Random Variables

Chapter Outline

2.1 Introduction
2.2 Random Variables
2.3 Probability Mass Function
2.4 Discrete Distribution Function
2.5 Probability Density Function
2.6 Continuous Distribution Function
2.7 Two-Dimensional Discrete Random Variables
2.8 Two-Dimensional Continuous Random Variables

2.1 INTRODUCTION

The outcomes of random experiments are, in general, abstract quantities or, in other words, most of the time they are not in any numerical form. However, the outcomes of a random experiment can be expressed in quantitative terms, in particular, by means of real numbers. Hence, a function can be defined that takes a definite real value corresponding to each outcome of an experiment. This gives a rationale for the concept of random variables about which probability statements can be made. In probability and statistics, a probability distribution assigns a probability to each measurable subset of the possible outcomes of a random experiment. Important and commonly encountered probability distributions include binomial distribution, Poisson distribution, and normal distribution.

2.2 RANDOM VARIABLES

A random variable X is a real-valued function of the elements of the sample space of a random experiment. In other words, a variable which takes the real values, depending on the outcome of a random experiment is called a *random variable*, e.g.,

(i) When a fair coin is tossed, $S = \{H, T\}$. If X is the random variable denoting the number of heads,

$X(H) = 1$ and $X(T) = 0$

Hence, the random variable X can take values 0 and 1.

(ii) When two fair coins are tossed, $S = \{HH, HT, TH, TT\}$. If X is the random variable denoting the number of heads,

$X(HH) = 2, X(HT) = 1, X(TH) = 1, X(TT) = 0.$

Hence, the random variable X can take values 0, 1, and 2.

(iii) When a fair die is tossed, $S = \{1, 2, 3, 4, 5, 6\}$.

If X is the random variable denoting the square of the number obtained,

$X(1) = 1, X(2) = 4, X(3) = 9, X(4) = 16, X(5) = 25, X(6) = 36$

Hence, the random variable X can take values 1, 4, 9, 16, 25, and 36.

Types of Random Variables

There are two types of random variables:

(i) Discrete random variables

(ii) Continuous random variables

Discrete Random Variables A random variable X is said to be discrete if it takes either finite or countably infinite values. Thus, a discrete random variable takes only isolated values, e.g.,

(i) Number of children in a family

(ii) Number of cars sold by different companies in a year

(iii) Number of days of rainfall in a city

(iv) Number of stars in the sky

(v) Profit made by an investor in a day

Continuous Random Variables A random variable X is said to be continuous if it takes any values in a given interval. Thus, a continuous random variable takes uncountably infinite values, e.g.,

(i) Height of a person in cm

(ii) Weight of a bag in kg

(iii) Temperature of a city in degree Celsius

(iv) Life of an electric bulb in hours

(v) Volume of a gas in cc.

Example 1

Identify the random variables as either discrete or continuous in each of the following cases:

(i) A page in a book can have at most 300 words

X = Number of misprints on a page

(ii) Number of students present in a class of 50 students

(iii) A player goes to the gymnasium regularly

X = Reduction in his weight in a month

(iv) Number of attempts required by a candidate to clear the IAS examination

(v) Height of a skyscraper

Solution

(i) X = Number of misprints on a page

The page may have no misprint or 1 misprint or 2 misprint … or 300 misprints. Thus, X takes values 0, 1, 2, …, 300. Hence, X is a discrete random variable.

(ii) Let X be the random variable denoting the number of students present in a class. X takes values 0, 1, 2, …, 50. Hence, X is a discrete random variable.

(iii) Reduction in weight cannot take isolated values 0, 1, 2, etc., but it takes any continuous value.

Hence, X is a continuous random variable.

(iv) Let X be a random variable denoting the number of attempts required by a candidate. Thus, X takes values 1, 2, 3, …. Hence, X is a discrete random variable.

(v) Since height can have any fractional value, it is a continuous random variable.

2.3 PROBABILITY MASS FUNCTION

Probability distribution of a random variable is the set of its possible values together with their respective probabilities. Let X be a discrete random variable which takes the values $x_1, x_2, \ldots x_n$. The probability of each possible outcome x_i is $p_i = p(x_i) = P(X = x_i)$ for $i = 1, 2, \ldots, n$. The number $p(x_i)$, $i = 1, 2, \ldots$ must satisfy the following conditions:

(i) $p(x_i) \geq 0$ for all values of i

(ii) $\sum_{i=1}^{\infty} p(x_i) = 1$

The function $p(x_i)$ is called the probability function or probability mass function of the random variable X. The set of pairs $\{x, p(x_i)\}$, $i = 1, 2, \ldots, n$ is called the probability distribution of the random variable which can be displayed in the form of a table as shown below:

$X = x_i$	x_1	x_2	x_3	$\dots x_i$	$\dots x_n$
$p(x_i) = P(X = x_i)$	$p(x_1)$	$p(x_2)$	$p(x_3)$	$\dots p(x_i)$	$\dots p(x_n)$

2.4 DISCRETE DISTRIBUTION FUNCTION

Let X be a discrete random variable which takes the values $x_1, x_2, \dots$ such that $x_1 < x_2 < \dots$ with probabilities $p(x_1), p(x_2) \dots$ such that $p(x_i) \ge 0$ for all values of i and $\sum_{i=1}^{x} p(x_i) = 1$.

The distribution function $F(x)$ of the discrete random variable X is defined by

$$F(x) = P(X \le x) = \sum_{i=1}^{x} p(x_i)$$

where x is any integer. The function $F(x)$ is also called the cumulative distribution function. The set of pairs $\{x_i, F(x)\}$, $i = 1, 2, \dots$ is called the cumulative probability distribution.

X	x_1	x_2	...
$F(x)$	$p(x_1)$	$p(x_1) + p(x_2)$	...

Example 1

A fair die is tossed once. If the random variable is getting an even number, find the probability distribution of X.

Solution

When a fair die is tossed,

$$S = \{1, 2, 3, 4, 5, 6\}$$

Let X be the random variable of getting an even number. Hence, X can take the values 0 and 1.

$$P(X = 0) = P(1, 3, 5) = \frac{3}{6} = \frac{1}{2}$$

$$P(X = 1) = P(2, 4, 6) = \frac{3}{6} = \frac{1}{2}$$

Hence, the probability distribution of X is

$X = x$	0	1
$P(X = 1)$	$\frac{1}{2}$	$\frac{1}{2}$

Also, $\sum P(X = x) = \frac{1}{2} + \frac{1}{2} = 1$

Example 2

Find the probability distribution of the number of heads when three coins are tossed.

Solution

When three coins are tossed,

$$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$$

Let X be the random variable of getting heads in tossing of three coins. Hence X can take the values 0, 1, 2, 3.

$$P(X = 0) = P(\text{no head}) = P(TTT) = \frac{1}{8}$$

$$P(X = 1) = P(\text{one head}) = P(HTT, THT, TTH) = \frac{3}{8}.$$

$$P(X = 2) = P(\text{two heads}) = P(HHT, THH, HTH) = \frac{3}{8}$$

$$P(X = 3) = P(\text{three heads}) = P(HHH) = \frac{1}{8}$$

Hence, the probability distribution of X is

$X = x$	0	1	2	3
$P(X = x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Also, $\sum P(X = x) = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8} = 1$

Example 3

State with reasons whether the following represent the probability mass function of a random variable:

(i)

$X = x$	0	1	2	3
$P(X = x)$	0.4	0.3	0.2	0.1

(ii)

$X = x$	0	1	2	3
$P(X = x)$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{4}$

(iii)

$X = x$	0	1	2	3
$P(X = x)$	$-\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{3}{4}$

Solution

(i) Here, $0 \le P(X = x) \le 1$ is satisfied for all values of X.

$$\begin{aligned}\sum P(X = x) &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)\\ &= 0.4 + 0.3 + 0.2 + 0.1\\ &= 1\end{aligned}$$

Since $\sum P(X = x) = 1$, it represents probability mass function.

(ii) Here, $0 \le P(X = x) \le 1$ is satisfied for all values of X.

$$\begin{aligned}\sum P(X = x) &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)\\ &= \frac{1}{2} + \frac{1}{3} + \frac{1}{6} + \frac{1}{4}\\ &= \frac{5}{4} > 1\end{aligned}$$

Since $\sum (P(X = x) > 1$, it does not represent a probability mass function.

(iii) Here, $0 \le P(X = x) \le 1$ is not satisfied for all the values of X as $P(X = 0) = -\frac{1}{2}$.

Hence, $P(X = x)$ does not represent a probability mass function.

Example 4

Verify whether the following functions can be regarded as the probability mass function for the given values of X:

(i) $P(X = x) = \frac{1}{5}$ *for* $x = 0, 1, 2, 3, 4$

$= 0$ *for otherwise*

(ii) $P(X = x) = \frac{x-2}{5}$ *for* $x = 1, 2, 3, 4, 5$

$= 0$ *for otherwise*

(iii) $P(X = x) = \frac{x^2}{30}$ *for* $x = 0, 1, 2, 3, 4$

$= 0$ *for otherwise*

Solution

(i) $P(X = 0) = P(X = 1) = P(X = 2) = P(X = 3) = P(X = 4) = \frac{1}{5}$

$P(X = x) \ge 0$ for all values of x

$$\begin{aligned}\sum P(X = x) &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)\\ &= \frac{1}{5} + \frac{1}{5} + \frac{1}{5} + \frac{1}{5} + \frac{1}{5}\\ &= 1\end{aligned}$$

Hence, $P(X = x)$ is a probability mass function.

(ii) $P(X = 1) = \frac{1-2}{5} = -\frac{1}{5} < 0$

Hence, $P(X = x)$ is not a probability mass function.

(iii) $P(X = 0) = 0$

$P(X = 1) = \frac{1}{30}$

$P(X = 2) = \frac{4}{30}$

$P(X = 3) = \frac{9}{30}$

$P(X = 4) = \frac{16}{30}$

$P(X = x) \ge 0$ for all values of x

$$\sum P(X=x) = P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4)$$
$$= 0 + \frac{1}{30} + \frac{4}{30} + \frac{9}{30} + \frac{16}{30}$$
$$= 1$$

Hence, $P(X=x)$ is a probability mass function.

Example 5

A random variable X has the probability mass function given by

X	1	2	3	4
$P(X=x)$	0.1	0.2	0.5	0.2

Find (i) $P(2 \le x < 4)$, (ii) $P(X > 2)$, (iii) P(X is odd), and (iv) P(X is even).

Solution

(i) $P(2 \le X < 4) = P(X=2) + P(X=3)$
$= 0.2 + 0.5$
$= 0.7$

(ii) $P(X>2) = P(X=3) + P(X=4)$
$= 0.5 + 0.2$
$= 0.7$

(iii) $P(X \text{ is odd}) = P(X=1) + P(X=3)$
$= 0.1 + 0.5$
$= 0.6$

(iv) $P(X \text{ is even}) = P(X=2) + P(X=4)$
$= 0.2 + 0.2$
$= 0.4$

Example 6

If the random variable X takes the value 1, 2, 3, and 4 such that $2P(X = 1) = 3P(X = 2) = P(X = 3) = 5P(X = 4)$. Find the probability distribution.

Solution

Let $2P(X=1) = 3P(X=2) = P(X=3) = 5P(X=4) = k$

$$P(X=1) = \frac{k}{2}$$

$$P(X=2) = \frac{k}{3}$$

$$P(X=3) = k$$

$$P(X=4) = \frac{k}{5}$$

Since $\sum(P(X=x) = 1$,

$$\frac{k}{2} + \frac{k}{3} + k + \frac{k}{5} = 1$$

$$k = \frac{30}{61}$$

Hence, the probability distribution is

X	1	2	3	4
$P(X=x)$	$\frac{15}{61}$	$\frac{10}{61}$	$\frac{30}{61}$	$\frac{6}{61}$

Example 7

A random variable X has the following probability distribution:

X	0	1	2	3	4	5	6	7
$P(X=x)$	a	$4a$	$3a$	$7a$	$8a$	$10a$	$6a$	$9a$

(i) Find the value of a.

(ii) Find $P(X < 3)$.

(iii) Find the smallest value of m for which $P(X \le m) \ge 0.6$.

Solution

(i) Since $P(X=x)$ is a probability distribution function,

$$\sum(P(X=x) = 1$$
$$P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) + P(X=5) + P(X=6) + P(X=7) = 1$$
$$a + 4a + 3a + 7a + 8a + 10a + 6a + 9a = 1$$
$$a = \frac{1}{48}$$

(ii) $P(X<3) = P(X=0) + P(X=1) + P(X=2)$
$= a + 4a + 3a$
$= 8a$
$= 8\left(\frac{1}{48}\right)$
$= \frac{1}{6}$

(iii) $P(X \le 4) = P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4)$

$$= a + 4a + 3a + 7a + 8a$$
$$= 23a$$
$$= 23\left(\frac{1}{48}\right)$$
$$= 0.575$$

$$P(X \le 5) = P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) + P(X=5)$$
$$= a + 4a + 3a + 7a + 8a + 10a$$
$$= 33a$$
$$= 33\left(\frac{1}{48}\right)$$
$$= 0.69$$

Hence, the smallest value of m for which $P(X \le m) \ge 0.6$ is 5.

Example 8

The probability mass function of a random variable X is zero except at the points X = 0, 1, 2. At these points, it has the values $P(X = 0) = 3c^3$, $P(X = 1) = 4c - 10c^2$, $P(X = 2) = 5c - 1$. *Find (i) c, (ii)* $P(X < 1)$, *(iii)* $P(1 < X \le 2)$, *and (iv)* $P(0 < X \le 2)$.

Solution

(i) Since $P(X = x)$ is a probability mass function,

$$\sum (P(X = x) = 1$$
$$P(X=0) + P(X=1) + P(X=2) = 1$$
$$3c^3 + 4c - 10c^2 + 5c - 1 = 1$$
$$3c^3 - 10c^2 + 9c - 2 = 0$$
$$(3c-1)(c-2)(c-1) = 0$$
$$c = \frac{1}{3},\ 2,\ 1$$

But $c < 1$, otherwise given probabilities will be greater than one or less than zero.

$$\therefore \quad c = \frac{1}{3}$$

Hence, the probability distribution is

X	0	1	2
$P(X = x)$	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{2}{3}$

(ii) $P(X < 1) = P(X = 0) = \frac{1}{9}$

(iii) $P(1 < X \le 2) = P(X = 2) = \frac{2}{3}$

(iv) $P(0 < X \le 2) = P(X = 1) + P(X = 2)$

$$= \frac{2}{9} + \frac{2}{3}$$
$$= \frac{8}{9}$$

Example 9

From a lot of 10 items containing 3 defectives, a sample of 4 items is drawn at random. Let the random variable X denote the number of defective items in the sample. Find the probability distribution of X.

Solution

The random variable X can take the value 0, 1, 2, or 3.

Total number of items = 10

Number of good items = 7

Number of defective items = 3

$$P(X=0) = P(\text{no defective}) = \frac{{}^7C_4}{{}^{10}C_4} = \frac{1}{6}$$

$$P(X=1) = P(\text{one defective and three good items}) = \frac{{}^3C_1\ {}^7C_3}{{}^{10}C_4} = \frac{1}{2}$$

$$P(X=2) = P(\text{two defectives and two good items}) = \frac{{}^3C_2\ {}^7C_2}{{}^{10}C_4} = \frac{3}{10}$$

$$P(X=3) = P(\text{three defectives and one good item}) = \frac{{}^3C_3 . {}^7C_1}{{}^{10}C_4} = \frac{1}{30}$$

Hence, the probability distribution of the random variable is

X	0	1	2	3
$P(X = x)$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{3}{10}$	$\frac{1}{30}$

Example 10

Construct the distribution function of the discrete random variable X whose probability distribution is as given below:

X	1	2	3	4	5	6	7
$P(X=x)$	0.1	0.15	0.25	0.2	0.15	0.1	0.05

Solution

Distribution function of X

X	$P(X=x)$	$F(x)$
1	0.1	0.1
2	0.15	0.25
3	0.25	0.5
4	0.2	0.7
5	0.15	0.85
6	0.1	0.95
7	0.05	1

Example 11

A random variable X has the probability function given below:

X	0	1	2
$P(X=x)$	k	$2k$	$3k$

Find (i) k, (ii) $P(X<2)$, $P(X\le 2)$, $P(0<X<2)$, *and (iii) the distribution function.*

Solution:

(i) Since $P(X=x)$ is a probability mass function,

$$\sum (P(X=x)=1$$
$$k+2k+3k=1$$
$$6k=1$$
$$k=\frac{1}{6}$$

Hence, the probability distribution is

X	0	1	2
$P(X=x)$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$

(ii) $P(X<2)=P(X=0)+P(X=1)=\frac{1}{6}+\frac{2}{6}=\frac{1}{2}$

$$P(X\le 2)=P(X=0)+P(X=1)+P(X=2)=\frac{1}{6}+\frac{2}{6}+\frac{3}{6}=1$$
$$P(0<X<2)=P(X=1)=\frac{1}{3}$$

(iii) Distribution function

X	$P(X=x)$	$F(x)$
0	$\frac{1}{6}$	$\frac{1}{6}$
1	$\frac{2}{6}$	$\frac{1}{2}$
2	$\frac{3}{6}$	1

Example 12

A random variable X takes the values $-3, -2, -1, 0, 1, 2, 3$, *such that* $P(X=0)=P(X>0)=P(X<0)$,
$P(X=-3)=P(X=-2)=P(X=-1)=P(X=1)=P(X=2)=P(X=3)$.
Obtain the probability distribution and the distribution function of X.

Solution

Let $P(X=0)=P(X>0)=P(X<0)=k_1$

Since $\sum P(X=x)=1$

$$k_1+k_1+k_1=1$$
$$\therefore \quad k_1=\frac{1}{3}$$
$$P(X=0)=P(X>0)=P(X<0)=\frac{1}{3}$$

Let $P(X=1)=P(X=2)=P(X=3)=k_2$

$$P(X>0)=P(X=1)+P(X=2)+P(X=3)$$
$$\frac{1}{3}=k_2+k_2+k_2$$
$$\therefore \quad k_2=\frac{1}{9}$$
$$P(X=1)=P(X=2)=P(X=3)=\frac{1}{9}$$

Similarly, $P(X=-3)=P(X=-2)=P(X=-1)=\frac{1}{9}$

Probability distribution and distribution function

X	$P(X=x)$	$F(x)$
-3	$\frac{1}{9}$	$\frac{1}{9}$
-2	$\frac{1}{9}$	$\frac{2}{9}$
-1	$\frac{1}{9}$	$\frac{3}{9}$
0	$\frac{1}{3}$	$\frac{6}{9}$
1	$\frac{1}{9}$	$\frac{7}{9}$
2	$\frac{1}{9}$	$\frac{8}{9}$
3	$\frac{1}{9}$	1

Example 13

A discrete random variable X has the following distribution function:

$$F(x)=\begin{cases} 0 & x<1 \\ \frac{1}{3} & 1\le x<4 \\ \frac{1}{2} & 4\le x<6 \\ \frac{5}{6} & 6\le x<10 \\ 1 & x\ge 10 \end{cases}$$

Find (i) $P(2<X\le 6)$, (ii) $P(X=5)$, (iii) $P(X=4)$, (iv) $P(X\le 6)$, and (v) $P(X=6)$.

Solution

(i) $P(2<X\le 6)=F(6)-F(2)=\frac{5}{6}-\frac{1}{3}=\frac{3}{6}=\frac{1}{2}$

(ii) $P(X=5)=P(X\le 5)-P(X<5)=F(5)-P(X<5)=\frac{1}{2}-\frac{1}{2}=0$

(iii) $P(X=4)=P(X\le 4)-P(X<4)=F(4)-P(X<4)=\frac{1}{2}-\frac{1}{3}=\frac{1}{6}$

(iv) $P(X\le 6)=F(6)=\frac{5}{6}$

(v) $P(X=6)=P(X\le 6)-P(X<6)=F(6)-P(X<6)=\frac{5}{6}-\frac{1}{2}=\frac{1}{3}$

EXERCISE 2.1

1. Verify whether the following functions can be considered as probability mass functions:

(i) $P(X=x)=\frac{x^2+1}{18}, x=0,1,2,3$ [Ans.: Yes]

(ii) $P(X=x)=\frac{x^2-2}{8}, x=1,2,3$ [Ans.: No]

(iii) $P(X=x)=\frac{2x+1}{18}, x=0,1,2,3$ [Ans.: No]

2. The probability mass function of a random variable X is

X	0	1	2	3	4	5	6
$P(X=x)$	k	$3k$	$5k$	$7k$	$9k$	$11k$	$13k$

Find $P(X<4)$ and $P(3<X\le 6)$.

$\left[\text{Ans.: } \frac{16}{49}, \frac{33}{49}\right]$

3. A random variable X has the following probability distribution:

X	1	2	3	4	5	6	7
$P(X=x)$	k	$2k$	$3k$	k^2	k^2+k	$2k^2$	$4k^2$

Find (i) k, (ii) $P(X<5)$, (iii) $P(X>5)$, and (iv) $P(0\le X\le 5)$

$\left[\text{Ans.: } \frac{1}{8} \text{ (ii) } \frac{49}{64} \text{ (iii) } \frac{3}{32} \text{ (iv) } \frac{29}{32}\right]$

4. A discrete random variable X has the following probability distribution:

X	-2	-1	0	1	2	3
$P(X = x)$	0.1	k	0.2	$2k$	0.3	$3k$

Find (i) k, (ii) $P(X \geq 2)$, and (iii) $P(-2 < X < 2)$.

$\left[\text{Ans.: } \frac{1}{15} \text{ (ii) } \frac{1}{2} \text{ (iii) } \frac{2}{5}\right]$

5. Given the following probability function of a discrete random variable X:

X	0	1	2	3	4	5	6	7
$P(X = x)$	0	c	$2c$	$2c$	$3c$	c^2	$2c^2$	$7c^2 + c$

Find (i) c, (ii) $P(X \geq 6)$, (iii) $P(X < 6)$, and (iv) Find k if $P(X \leq k) > \frac{1}{2}$, where k is a positive integer.

[**Ans.:** (i) 0.1 (ii) 0.19 (iii) 0.81 (iv) 4]

6. A random variable X assumes four values with probabilities $\frac{1+3x}{4}, \frac{1-x}{4}, \frac{1+2x}{4}$ and $\frac{1-4x}{4}$. For what value of x do these values represent the probability distribution of X?

$\left[\text{Ans.: } -\frac{1}{3} \leq X \leq \frac{1}{4}\right]$

7. Let X denote the number of heads in a single toss of 4 fair coins. Determine (i) $P(X < 2)$, and (ii) $P(1 < X \leq 3)$.

$\left[\text{Ans.: (i) } \frac{5}{16} \text{ (ii) } \frac{5}{8}\right]$

8. If 3 cars are selected from a lot of 6 cars containing 2 defective cars, find the probability distribution of the number of defective cars.

Ans.:

X	0	1	2
$P(X = x)$	$\frac{1}{5}$	$\frac{3}{5}$	$\frac{2}{5}$

9. Five defective bolts are accidentally mixed with 20 good ones. Find the probability distribution of the number of defective bolts, if four bolts are drawn at random from this lot.

Ans.:

X	0	1	2	3	4
$P(X = x)$	$\frac{969}{2530}$	$\frac{1140}{2530}$	$\frac{380}{2530}$	$\frac{40}{2530}$	$\frac{1}{2530}$

10. Two dice are rolled at once. Find the probability distribution of the sum of the numbers on them.

Ans.:

X	2	3	4	5	6	7	8	9	10	11	12
$P(X = x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

11. A random variable X takes three values 0, 1, and 2 with probabilities $\frac{1}{3}, \frac{1}{6}$, and $\frac{1}{2}$ respectively. Obtain the distribution function of X.

$\left[\text{Ans.: } F(0) = \frac{1}{3}, F(1) = \frac{1}{2}, F(2) = 1\right]$

12. A random variable X has the following probability function:

X	0	1	2	3	4
$P(X = x)$	k	$3k$	$5k$	$7k$	$9k$

Find (i) the value of k, (ii) $P(X < 3)$, $P(X \geq 3)$, $P(0 < X < 4)$, and (iii) distribution function of X.

$$\left[\begin{array}{l} \text{Ans.: (i) } \frac{1}{25}, \text{ (ii) } \frac{9}{25}, \frac{16}{25}, \frac{3}{5} \\ \qquad \text{(iii) } F(0) = \frac{1}{25}, F(1) = \frac{4}{25}. F(2) = \frac{9}{25}, F(3) = \frac{16}{25}, F(4) = 1 \end{array}\right]$$

13. A random variable X has the probability function

X	-2	-1	0	1	2	3
$P(X = x)$	0.1	k	0.2	$2k$	0.3	k

Find (i) k, (ii) $P(X \leq 1)$, (iii) $P(-2 < X < 1)$, and (iv) obtain the distribution function of X.

[**Ans.:** (i) 0.1 (ii) 0.6 (iii) 0.3]

14. The following is the distribution function $F(x)$ of a discrete random variable X:

X	-3	-2	-1	0	1	2	3
$P(X = x)$	0.08	0.2	0.4	0.65	0.8	0.9	1

Find (i) the probability distribution of X, (ii) $P(-2 \le X \le 1)$, and (iii) $P(X \ge 1)$.

Ans.: (i)

X	–3	–2	–1	0	1	2	3
$P(X = x)$	0.08	0.12	0.2	0.25	0.15	0.1	0.1

(ii) 0.72 (ii) 0.35.

2.5 PROBABILITY DENSITY FUNCTION

Let X be a continuous random variable such that the probability of the variable X falling in the small interval $x-\frac{1}{2}\mathrm{d}x$ to $x+\frac{1}{2}\mathrm{d}x$ is $f(x)\,\mathrm{d}x$, i.e.,

$$P\left(x-\frac{1}{2}\mathrm{d}x \le X \le x+\frac{1}{2}\mathrm{d}x\right)=f(x)\mathrm{d}x$$

The function $f(x)$ is called the probability density function of the random variable X and the continuous curve $y=f(x)$ is called the probability curve.

Properties of Probability Density Function

(i) $f(x) \ge 0, \qquad -\infty < x < \infty$

(ii) $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$

(iii) $P(a<x<b)=\int_a^b f(x)\,\mathrm{d}x$

2.6 CONTINUOUS DISTRIBUTION FUNCTION

If X is a continuous random variable having the probability density function $f(x)$ then the function

$$F(x)=P(X \le x)=\int_{-\infty}^{x} f(x)\mathrm{d}x, \quad -\infty<x<\infty$$

is called the distribution function or cumulative distribution function of the random variable X.

Properties of Cumulative Distribution Function

(i) $F(-\infty)=0$

(ii) $F(\infty)=1$

(iii) $0 \le F(x) \le 1, \quad -\infty<x<\infty$

(iv) $P(a<X<b)=F(b)-F(a)$

(v) $F'(x)=\frac{\mathrm{d}}{\mathrm{d}x}F(x)=f(x), \quad f(x) \ge 0$

Example 1

Show that the function $f(x)$ defined by

$$f(x)=\frac{1}{7} \qquad 1<x<8$$
$$=0 \qquad \textit{otherwise}$$

is a probability density function for a random variable. Hence, find $P(3<X<10)$.

Solution

$$f(x) \ge 0 \quad \text{in} \quad 1<x<8$$

$$\begin{aligned}
\int_{-\infty}^{\infty} f(x)\mathrm{d}x &= \int_{-\infty}^{1} f(x)\mathrm{d}x+\int_{1}^{8} f(x)\mathrm{d}x+\int_{8}^{\infty} f(x)\mathrm{d}x \\
&=0+\int_1^8 \frac{1}{7}\mathrm{d}x+0 \\
&=\frac{1}{7}|x|_1^8 \\
&=\frac{1}{7}(8-1) \\
&=1
\end{aligned}$$

Hence, $f(x)$ is a probability density function.

$$\begin{aligned}
P(3<X<10) &= \int_3^{10} f(x)\mathrm{d}x \\
&=\int_3^8 f(x)\mathrm{d}x+\int_8^{10} f(x)\mathrm{d}x \\
&=\int_3^8 \frac{1}{7}\mathrm{d}x+0 \\
&=\frac{1}{7}|x|_3^8 \\
&=\frac{1}{7}(8-3) \\
&=\frac{5}{7}
\end{aligned}$$

Example 2

Is the function f(x) defined by

$$f(x)=e^{-x} \quad x\ge 0$$
$$=0 \quad x<0$$

is a probability density function. If so, find the probability that the variate having this density falls in the interval (1, 2).

Solution

$$f(x)\ge 0 \quad \text{in} \quad (0,\infty)$$

$$\int_{-\infty}^{\infty} f(x)\,dx=\int_{-\infty}^{0} f(x)\,dx+\int_{0}^{\infty} f(x)\,dx$$
$$=0+\int_{0}^{\infty} e^{-x}\,dx$$
$$=\left|-e^{-x}\right|_{0}^{\infty}$$
$$=-e^{-\infty}+1$$
$$=1$$

Hence, $f(x)$ is a probability density function.

$$P(1\le X\le 2)=\int_{1}^{2} f(x)\,dx$$
$$=\int_{1}^{2} e^{-x}\,dx$$
$$=\left|-e^{-x}\right|_{1}^{2}$$
$$=-e^{-2}+e^{-1}$$
$$=0.233$$

Example 3

If a random variable has the probability density function f(x) as

$$f(x)=2e^{-2x} \quad x>0$$
$$=0 \quad x\le 0$$

Find the probabilities that it will take on a value (i) between 1 *and* 3, *and (ii) greater than 0.5.*

Solution

(i) Probability that the variable will take a value between 1 and 3

$$P(1<X<3)=\int_{1}^{3} f(x)\,dx$$
$$=\int_{1}^{3} 2e^{-2x}\,dx$$
$$=2\left|\frac{e^{-2x}}{-2}\right|_{1}^{3}$$
$$=-(e^{-6}-e^{-2})$$
$$=e^{-2}-e^{-6}$$

(ii) Probability that the variable will take a value greater than 0.5

$$P(X>0.5)=\int_{0.5}^{\infty} f(x)\,dx$$
$$=\int_{0.5}^{\infty} 2e^{-2x}\,dx$$
$$=2\left|\frac{e^{-2x}}{-2}\right|_{0.5}^{\infty}$$
$$=-(e^{-\infty}-e^{-1})$$
$$=e^{-1}$$

Example 4

Find the constant k such that the function

$$f(x)=kx^2 \quad 0<x<3$$
$$=0 \quad \text{otherwise}$$

is a probability density function and compute (i) $P(1<x<2)$, *(ii)* $P(X<2)$, *and (iii)* $P(X\ge 2)$.

Solution

Since $f(x)$ is a probability density function,

$$\int_{-\infty}^{\infty} f(x)\,dx=1$$

$$\int_{-\infty}^{0} f(x)\,dx+\int_{0}^{3} f(x)\,dx+\int_{3}^{\infty} f(x)\,dx=1$$

$$0+\int_{0}^{3} kx^2\,\mathrm{d}x+0=1$$

$$k\left|\frac{x^3}{3}\right|_0^3=1$$

$$\frac{k}{3}(27-0)=1$$

$$9k=1$$

$$k=\frac{1}{9}$$

Hence, $f(x)=\frac{1}{9}x^2 \quad 0<x<3$

$=0$ otherwise.

(i) $P(1<X<2)=\int_1^2 f(x)\,\mathrm{d}x$

$$=\int_1^2 \frac{1}{9}x^2\,\mathrm{d}x$$

$$=\frac{1}{9}\left|\frac{x^3}{3}\right|_1^2$$

$$=\frac{1}{27}(8-1)$$

$$=\frac{7}{27}$$

(ii) $P(X<2)=\int_{-\infty}^{2} f(x)\,\mathrm{d}x$

$$=\int_{-\infty}^{0} f(x)\,\mathrm{d}x+\int_0^2 f(x)\,\mathrm{d}x$$

$$=0+\int_0^2 \frac{1}{9}x^2\,\mathrm{d}x$$

$$=\frac{1}{9}\int_0^2 x^2\,\mathrm{d}x$$

$$=\frac{1}{9}\left|\frac{x^3}{3}\right|_0^2$$

$$=\frac{1}{27}(8-0)$$

$$=\frac{8}{27}$$

(iii) $P(X\ge 2)=1-P(X<2)$

$$=1-\frac{8}{27}$$

$$=\frac{19}{27}$$

Example 5

If the probability density function of a random variable is given by

$$f(x)=k(1-x^2) \quad 0<x<1$$
$$=0 \quad \textit{otherwise}$$

Find the value of k and the probabilities that a random variable having this probability density will take on a value (i) between 0.1 *and* 0.2*, and (ii) greater than* 0.5.

Solution

Since $f(x)$ is a probability density function,

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x=1$$

$$\int_{-\infty}^{0} f(x)\,\mathrm{d}x+\int_{0}^{1} f(x)\,\mathrm{d}x+\int_{1}^{\infty} f(x)\,\mathrm{d}x=1$$

$$0+\int_{0}^{1} k(1-x^2)\,\mathrm{d}x+0=1$$

$$k\left|x-\frac{x^3}{3}\right|_0^1=1$$

$$k\left(1-\frac{1}{3}\right)=1$$

$$k=\frac{3}{2}$$

Hence, $f(x)=\frac{3}{2}(1-x^2) \quad 0<x<1$

$=0$ otherwise

(i) Probability that the variable will take on a value between 0.1 and 0.2

$$P(0.1<X<0.2)=\int_{0.1}^{0.2} f(x)\,\mathrm{d}x$$

$$=\int_{0.1}^{0.2} \frac{3}{2}(1-x^2)\,\mathrm{d}x$$

$$= \frac{3}{2}\left|x - \frac{x^3}{3}\right|_{0.1}^{0.2}$$

$$= \frac{3}{2}\left[\left(0.2 - \frac{0.008}{3}\right) - \left(0.1 - \frac{0.001}{3}\right)\right]$$

$$= 0.1465$$

(ii) Probability that the variable will take on a value greater than 0.5

$$P(X > 0.5) = \int_{0.5}^{\infty} f(x)\,dx$$

$$= \int_{0.5}^{1} f(x)\,dx + \int_{1}^{\infty} f(x)\,dx$$

$$= \int_{0.5}^{1} \frac{3}{2}(1 - x^2)\,dx + 0$$

$$= \frac{3}{2}\left|x - \frac{x^3}{3}\right|_{0.5}^{1}$$

$$= \frac{3}{2}\left[\left(1 - \frac{1}{3}\right) - \left(0.5 - \frac{0.125}{3}\right)\right]$$

$$= 0.3125$$

Example 6

If X is a continuous random variable with pdf

$$f(x) = x^2 \quad 0 \le x \le 1$$
$$= 0 \quad \textit{otherwise}$$

If $P(a \le X \le 1) = \frac{19}{81}$, *find the value of a.*

Solution

$$P(a \le X \le 1) = \frac{19}{81}$$

$$\int_a^1 f(x)\,dx = \frac{19}{81}$$

$$\int_a^1 x^2\,dx = \frac{19}{81}$$

$$\left|\frac{x^3}{3}\right|_a^1 = \frac{19}{81}$$

$$\frac{1}{3}(1 - a) = \frac{19}{81}$$

$$1 - a = \frac{19}{27}$$

$$a = \frac{46}{27}$$

Example 7

Let X be a continuous random variable with pdf

$$f(x) = kx\,(1 - x),\ 0 \le x \le 1$$

Find k and determine a number b such that $P(X \le b) = P(X \ge b)$.

Solution

Since $f(x)$ is a probability density function,

$$\int_{-\infty}^{\infty} f(x) = 1$$

$$\int_{-\infty}^{0} f(x)\,dx + \int_{0}^{1} f(x)\,dx + \int_{1}^{\infty} f(x)\,dx = 1$$

$$0 + \int_0^1 kx(1 - x)\,dx + 0 = 1$$

$$k\int_0^1 (x - x^2)\,dx = 1$$

$$k\left|\frac{x^2}{2} - \frac{x^3}{3}\right|_0^1 = 1$$

$$k\left[\left(\frac{1}{2} - \frac{1}{3}\right) - (0 - 0)\right] = 1$$

$$k\left(\frac{1}{6}\right) = 1$$

$$k = 6$$

Hence, $f(x) = 6(x - x^2) \quad 0 \le x \le 1$

Since total probability is 1 and $P(X \le b) = P(X \ge b)$,

$$P(X \le b) = \frac{1}{2}$$

$$\int_0^b f(x)\,\mathrm{d}x = \frac{1}{2}$$

$$6\int_0^b (x - x^2)\,\mathrm{d}x = \frac{1}{2}$$

$$6\left|\frac{x^2}{2} - \frac{x^3}{3}\right|_0^b = \frac{1}{2}$$

$$\frac{b^2}{2} - \frac{b^3}{3} = \frac{1}{12}$$

$$6b^2 - 4b^3 = 1$$

$$4b^3 - 6b^2 + 1 = 0$$

$$(2b-1)(2b^2 - 2b - 1) = 0$$

$$b = \frac{1}{2} \text{ or } b = \frac{1 \pm \sqrt{3}}{2}$$

b lies in (0, 1).

$\therefore \quad b = \frac{1}{2}$

Example 8

The length of time (in minutes) that a certain lady speaks on the telephone is found to be a random phenomenon, with a probability function specified by the function

$$f(x) = A\,e^{-\frac{x}{5}} \quad x \ge 0$$
$$= 0 \quad \textit{otherwise}$$

(i) Find the value of A that makes f(x) a probability density function.
(ii) What is the probability that the number of minutes that she will take over the phone is more than 10 *minutes?*

Solution

(i) For $f(x)$ to be a probability density function,

$$\int_{-\infty}^{\infty} f(x)\,x = 1$$

$$\int_{-\infty}^{0} f(x)\,\mathrm{d}x + \int_{0}^{\infty} f(x)\,\mathrm{d}x = 1$$

$$0 + \int_0^{\infty} A e^{-\frac{x}{5}}\,\mathrm{d}x = 1$$

$$A\left|\frac{e^{-\frac{x}{5}}}{-\frac{1}{5}}\right|_0^{\infty} = 1$$

$$-5A(e^{-\infty} - e^{-0}) = 1$$

$$-5A(0-1) = 1$$

$$5A = 1$$

$$A = \frac{1}{5}$$

Hence, $f(x) = \frac{1}{5} e^{-\frac{x}{5}} \quad x \ge 0.$
$\qquad = 0 \qquad$ otherwise

(ii) $P(X > 10) = \int_{10}^{\infty} f(x)\,\mathrm{d}x$

$$= \int_{10}^{\infty} \frac{1}{5} e^{-\frac{x}{5}}\,\mathrm{d}x$$

$$= \frac{1}{5}\left|\frac{e^{-\frac{x}{5}}}{-\frac{1}{5}}\right|_{10}^{\infty}$$

$$= -\left(e^{-\infty} - e^{-2}\right)$$

$$= -\left(0 - e^{-2}\right)$$

$$= e^{\frac{1}{2}}$$

Example 9

A continuous random variable X has a pdf $f(x)^2 = 3x^2, 0 \le x \le 1$. *Find a and b such that*

(i) $P(X \le a) = P(X > a)$ and

(ii) $P(X > b) = 0.05$

Solution

Since total probability is 1 and $P(X \le a) = P(X > a)$,

$$P(X \le a) = \frac{1}{2}$$

$$\int_0^a f(x)\,dx = \frac{1}{2}$$

$$\int_0^a 3x^2\,dx = \frac{1}{2}$$

$$3\left|\frac{x^3}{3}\right|_0^a = \frac{1}{2}$$

$$a^3 = \frac{1}{2}$$

$$a = \left(\frac{1}{2}\right)^{\frac{1}{3}}$$

$$P(X > b) = 0.05$$

$$\int_b^1 f(x)\,dx = 0.05$$

$$\int_b^1 3x^2\,dx = 0.05$$

$$3\left|\frac{x^3}{3}\right|_b^1 = 0.05$$

$$1 - b^3 = 0.05$$

$$b^3 = \frac{19}{20}$$

$$b = \left(\frac{19}{20}\right)^{\frac{1}{3}}$$

Example 10

Let the continuous random variable X have the probability density function

$$f(x) = \frac{2}{x^3} \qquad 1 < x < \infty$$

$$= 0 \qquad \textit{otherwise}$$

Find F(x).

Solution

$$F(x) = \int_{-\infty}^{x} f(x)\,dx$$

$$= \int_{-\infty}^{1} f(x)\,dx + \int_{1}^{x} f(x)\,dx$$

$$= 0 + \int_1^x \frac{2}{x^3}\,dx$$

$$= 2\left|\frac{x^{-2}}{-2}\right|_1^x$$

$$= -\left|\frac{1}{x^2}\right|_1^x$$

$$= -\left(\frac{1}{x^2} - 1\right)$$

$$= 1 - \frac{1}{x^2}$$

Hence, $F(x) = 1 - \frac{1}{x^2} \qquad 1 < x < \infty$

$= 0 \qquad$ otherwise

Example 11

Verify that the function F(x) is a distribution function.

$$F(x) = 0 \qquad x < 0$$

$$= 1 - e^{-\frac{x}{4}} \qquad x \ge 0$$

Also, find the probabilities $P(X \le 4), P(X \ge 8), P(4 \le X \le 8)$.

Solution

For the function $F(x)$,

(i) $F(-\infty) = 0$

(ii) $F(\infty) = 1 - e^{-\infty} = 1 - 0 = 1$

(iii) $0 \le F(x) \le 1 \qquad -\infty < x < \infty$

If $f(x)$ is the corresponding probability density function,

$$f(x) = F'(x) = 0 \qquad x < 0$$

$$= \frac{1}{4}e^{-\frac{x}{4}} \qquad x \ge 0$$

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = \int_{-\infty}^{0} f(x)\mathrm{d}x + \int_{0}^{\infty} f(x)\mathrm{d}x$$

$$= 0 + \int_{0}^{\infty} \frac{1}{4} e^{-\frac{x}{4}}\mathrm{d}x$$

$$= \frac{1}{4}\left|\frac{e^{-\frac{x}{4}}}{-\frac{1}{4}}\right|_{0}^{\infty}$$

$$= -\left|e^{-\frac{x}{4}}\right|_{0}^{\infty}$$

$$= -(0-1)$$

$$= 1$$

Hence, $F(x)$ is a distribution function.

$$P(X \le 4) = F(4)$$

$$= 1 - e^{-1}$$

$$= 1 - \frac{1}{e}$$

$$= \frac{e-1}{e}$$

$$P(X \ge 8) = 1 - P(X \le 8)$$

$$= 1 - F(8)$$

$$= 1 - (1 - e^{-2})$$

$$= e^{-2}$$

$$= \frac{1}{e^2}$$

$$P(4 \le X \le 8) = F(8) - F(4)$$

$$= (1 - e^{-2}) - (1 - e^{-1})$$

$$= e^{-1} - e^{-2}$$

$$= \frac{1}{e} - \frac{1}{e^2}$$

$$= \frac{e-1}{e^2}$$

Example 12

The troubleshooting capacity of an IC chip in a circuit is a random variable X whose distribution function is given by

$$F(x) = 0 \qquad x \le 3$$

$$= 1 - \frac{9}{x^2} \qquad x > 3$$

where x denotes the number of years. Find the probability that the IC chip will work properly (i) less than 8 years, (i) beyond 8 years, (iii) between 5 to 7 years, and (iv) anywhere from 2 to 5 years.

Solution

(i) $P(X \le 8) = F(8)$

$$= 1 - \frac{9}{8^2}$$

$$= 0.8594$$

(ii) $P(X > 8) = 1 - P(X \le 8)$

$$= 1 - F(8)$$

$$= 1 - 0.8594$$

$$= 0.1406$$

(iii) $P(5 \le X \le 7) = F(7) - F(5)$

$$= \left(1 - \frac{9}{7^2}\right) - \left(1 - \frac{9}{5^2}\right)$$

$$= 0.1763$$

(iv) $P(2 \le X \le 5) = F(5) - F(2)$

$$= \left(1 - \frac{9}{5^2}\right) - 0$$

$$= 0.64$$

Example 13

The probability density function of a continuous random variable X is given by

$$f(x) = \begin{cases} ax & 0 \le x \le 1 \\ a & 1 \le x \le 2 \\ 3a - ax & 2 \le x \le 3 \\ 0 & \text{otherwise} \end{cases}$$

(i) Find the value of a, and (ii) find the cdf of X.

Solution

(i) Since $f(x)$ is a probability density function,

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$$

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x + \int_{0}^{1} f(x)\,\mathrm{d}x + \int_{1}^{2} f(x)\,\mathrm{d}x + \int_{2}^{3} f(x)\,\mathrm{d}x = 1$$

$$0 + \int_{0}^{1} ax\,\mathrm{d}x + \int_{1}^{2} a\,\mathrm{d}x + \int_{2}^{3} (3a - ax)\,\mathrm{d}x = 1$$

$$a\left|\frac{x^2}{2}\right|_0^1 + a\left|x\right|_1^2 + \left|3ax - \frac{ax^2}{2}\right|_2^3 = 1$$

$$a\left(\frac{1}{2} - 0\right) + a(2-1) + \left[\left(9a - \frac{9a}{2}\right) - (6a - 2a)\right] = 1$$

$$\frac{1}{2}a + a + \frac{9a}{2} - 4a = 1$$

$$2a = 1$$

$$a = \frac{1}{2}$$

(ii) $F(x) = \int_{-\infty}^{x} f(x)\,\mathrm{d}x$

For $0 \le x \le 1$,

$$F(x) = \int_{-\infty}^{0} f(x)\,\mathrm{d}x + \int_{0}^{x} f(x)\,\mathrm{d}x$$

$$= 0 + \int_{0}^{x} ax\,\mathrm{d}x$$

$$= a\left|\frac{x^2}{2}\right|_0^x$$

$$= \frac{ax^2}{2}$$

For $1 \le x \le 2$,

$$F(x) = \int_{-\infty}^{0} f(x)\,\mathrm{d}x + \int_{0}^{1} f(x)\,\mathrm{d}x + \int_{1}^{x} f(x)\,\mathrm{d}x$$

$$= 0 + \int_{0}^{1} ax\,\mathrm{d}x + \int_{1}^{x} a\,\mathrm{d}x$$

$$= a\left|\frac{x^2}{2}\right|_0^1 + a\left|x\right|_1^x$$

$$= a\left(\frac{1}{2} - 0\right) + a(x-1)$$

$$= \frac{a}{2} + ax - a$$

$$= ax - \frac{a}{2}$$

For $2 \le x \le 3$,

$$F(x) = \int_{-\infty}^{0} f(x)\,\mathrm{d}x + \int_{0}^{1} f(x)\,\mathrm{d}x + \int_{1}^{2} f(x)\,\mathrm{d}x + \int_{2}^{x} f(x)\,\mathrm{d}x$$

$$= 0 + \int_{0}^{1} ax\,\mathrm{d}x + \int_{1}^{2} a\,\mathrm{d}x + \int_{2}^{x} (3a - ax)\,\mathrm{d}x$$

$$= a\left|\frac{x^2}{2}\right|_0^1 + a\left|x\right|_1^2 + \left|3ax - \frac{ax^2}{2}\right|_2^x$$

$$= a\left(\frac{1}{2} - 0\right) + a(2-1) + \left[\left(3ax - \frac{ax^2}{2}\right) - (6a - 2a)\right]$$

$$= \frac{a}{2} + a + 3ax - \frac{ax^2}{2} - 4a$$

$$= 3ax - \frac{ax^2}{2} - \frac{5a}{2}$$

Hence,

$$\begin{aligned} F(x) &= \frac{ax^2}{2} && 0 \le x \le 1 \\ &= ax - \frac{a}{2} && 1 \le x \le 2 \\ &= 3ax - \frac{ax^2}{2} - \frac{5a}{2} && 2 \le x \le 3 \end{aligned}$$

Example 14

The pdf of a continuous random variable X is

$$f(x) = \frac{1}{2}e^{-|x|}$$

Find cdf F(x).

Solution

$$f(x)=\frac{1}{2}e^{x} \qquad -\infty<x<0$$

$$=\frac{1}{2}e^{-x} \qquad 0<x<\infty$$

$$F(x)=\int_{-\infty}^{x} f(x)\,dx$$

For $x\le 0$,

$$F(x)=\int_{-\infty}^{x}\frac{1}{2}e^{x}dx$$

$$=\frac{1}{2}\left|e^{x}\right|_{-\infty}^{x}$$

$$=\frac{1}{2}(e^{x}-e^{-\infty})$$

$$=\frac{1}{2}e^{x}$$

For $x>0$,

$$F(x)=\int_{-\infty}^{0} f(x)\,dx+\int_{0}^{x} f(x)\,dx$$

$$=\int_{-\infty}^{0}\frac{1}{2}e^{x}dx+\int_{0}^{x}\frac{1}{2}e^{-x}\,dx$$

$$=\frac{1}{2}\left|e^{x}\right|_{-\infty}^{0}+\frac{1}{2}\left|-e^{-x}\right|_{0}^{x}$$

$$=\frac{1}{2}(1-e^{-\infty})+\frac{1}{2}(-e^{-x}+e^{0})$$

$$=\frac{1}{2}-\frac{1}{2}e^{-x}+\frac{1}{2}$$

$$=1-\frac{1}{2}e^{-x}$$

Hence, $F(x)=\frac{1}{2}e^{x} \qquad x\le 0$

$$=1-\frac{1}{2}e^{-x} \qquad x>0$$

Example 15

Find the value of k and the distribution function F(x) given the probability density function of a random variable X as

$$f(x)=\frac{k}{x^{2}+1} \qquad -\infty<x<\infty$$

Solution

Since $f(x)$ is the probability density function,

$$\int_{-\infty}^{\infty} f(x)\,dx=1$$

$$\int_{-\infty}^{\infty}\frac{k}{x^{2}+1}dx=1$$

$$k\int_{-\infty}^{\infty}\frac{1}{x^{2}+1}dx=1$$

$$k\left|\tan^{-1}x\right|_{-\infty}^{\infty}=1$$

$$k\left[\tan^{-1}\infty-\tan^{-1}(-\infty)\right]=1$$

$$k\left[\frac{\pi}{2}-\left(-\frac{\pi}{2}\right)\right]=1$$

$$k\pi=1$$

$$k=\frac{1}{\pi}$$

Hence, $f(x)=\frac{1}{\pi}\frac{1}{x^{2}+1} \qquad -\infty<x<\infty$

$$F(x)=\int_{-\infty}^{x} f(x)\,dx$$

$$=\frac{1}{\pi}\int_{-\infty}^{x}\frac{1}{x^{2}+1}dx$$

$$=\frac{1}{\pi}\left|\tan^{-1}x\right|_{-\infty}^{x}$$

$$=\frac{1}{\pi}\left[\tan^{-1}x-\tan^{-1}(-\infty)\right]$$

$$=\frac{1}{\pi}\left(\tan^{-1}x+\frac{\pi}{2}\right)$$

Example 16

Find the constant k such that

$$f(x) = kx^2 \qquad 0 < x < 3$$
$$= 0 \qquad \text{otherwise}$$

is a probability function. Also, find the distribution function F(x) and $P(1 < X \le 2)$.

Solution

Since $f(x)$ is probability density function,

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$\int_{-\infty}^{0} f(x)\,dx + \int_{0}^{3} f(x)\,dx + \int_{3}^{\infty} f(x)\,dx = 1$$

$$0 + \int_{0}^{3} kx^2\,dx + 0 = 1$$

$$k\left|\frac{x^3}{3}\right|_0^3 = 1$$

$$k(9-0) = 1$$

$$k = \frac{1}{9}$$

Hence, $f(x) = \frac{1}{9}x^2 \qquad 0 < x < 3$

$$= 0 \qquad \text{otherwise}$$

$$F(x) = \int_{-\infty}^{x} f(x)\,dx$$

$$= \int_{-\infty}^{0} f(x)\,dx + \int_{0}^{x} f(x)\,dx$$

$$= 0 + \int_0^x \frac{1}{9}x^2\,dx$$

$$= \frac{1}{9}\left|\frac{x^3}{3}\right|_0^x$$

$$= \frac{1}{27}x^3$$

Hence, $F(x) = \frac{1}{27}x^3 \qquad 0 < x < 3$

$$= 0 \qquad \text{otherwise}$$

$$P(1 < x \le 2) = \int_{1}^{2} f(x)\,dx$$

$$= \int_{1}^{2} \frac{1}{9}x^2\,dx$$

$$= \frac{1}{9}\left|\frac{x^3}{3}\right|_1^2$$

$$= \frac{1}{27}(8-1)$$

$$= \frac{7}{27}$$

EXERCISE 2.2

1. Verify whether the following functions are probability density functions:

   (i) $f(x) = k\,e^{-kx} \qquad x \ge 0, k > 0$

   (ii) $f(x) = \frac{1}{2}e^{-|x|} \qquad -\infty < x < \infty$

   (iii) $f(x) = \frac{2}{9}x\left(2 - \frac{x}{2}\right) \qquad 0 \le x \le 3$

   [**Ans.:** (i) Yes (ii) Yes (iii) Yes]

2. Find the value of k if the following are probability density functions:

   (i) $f(x) = k(1+x) \qquad 2 \le x \le 5$

   (ii) $f(x) = k(x - x^2) \qquad 0 \le x \le 1$

   (iii) $f(x) = kx\,e^{-4x^2} \qquad 0 \le x \le \infty$

   (iv) $f(x) = kx\,e^{-\frac{x^2}{4}} \qquad 0 \le x \le \infty$

   [**Ans.:** (i) $\frac{2}{27}$ (ii) 6 (iii) 8 (iv) $\frac{1}{2}$]

3. A function is defined as

$$f(x) = \begin{cases} 0 & x < 2 \\ \dfrac{2x+3}{18} & 2 \le x \le 4 \\ 0 & x > 4 \end{cases}$$

Show that $f(x)$ is a probability density function and find $P(2 < X < 3)$.

$\left[\textbf{Ans.: } \dfrac{4}{9}\right]$

4. Let X be a continuous random variable with probability distribution

$$f(x) = \begin{cases} \dfrac{x}{6} + k & 0 \le x \le 3 \\ 0 & \text{otherwise} \end{cases}$$

Find k, and $P(1 \le X \le 2)$.

$\left[\textbf{Ans.: } 1, \dfrac{1}{3}\right]$

5. Find the value of k such that $f(x)$ is a probability density function. Find also, $P(X \le 1.5)$.

$$f(x) = \begin{cases} kx & 0 \le x \le 1 \\ k & 1 \le x \le 2 \\ k(3-x) & 2 \le x \le 3 \end{cases}$$

$\left[\textbf{Ans.: } \dfrac{1}{2}, \dfrac{1}{2}\right]$

6. If X is a continuous random variable whose probability density function is given by

$$\begin{aligned} f(x) &= k(4x - 2x^2) && 0 < x < 2 \\ &= 0 && \text{otherwise} \end{aligned}$$

Find (i) the value of k, and (ii) $P(X > 1)$.

$\left[\textbf{Ans.: } \text{(i) } \dfrac{3}{8} \text{ (ii) } \dfrac{1}{2}\right]$

7. If a random variable has the probability density function

$$\begin{aligned} f(x) &= k(x^2 - 1) && -1 \le x \le 3 \\ &= 0 && \text{otherwise} \end{aligned}$$

Find (i) the value of k, and (ii) $P\left(\dfrac{1}{2} \le X \le \dfrac{5}{2}\right)$.

$\left[\textbf{Ans.: } \text{(i) } \dfrac{3}{28} \text{ (ii) } \dfrac{19}{56}\right]$

8. The probability density function is

$$\begin{aligned} f(x) &= k(3x^2 - 1) && -1 \le x \le 2 \\ &= 0 && \text{otherwise} \end{aligned}$$

Find (i) the value of k, and (ii) $P(-1 \le X \le 0)$.

$\left[\textbf{Ans.: } \text{(i) } \dfrac{1}{6} \text{ (ii) } 0\right]$

9. Is the function defined by

$$\begin{aligned} f(x) &= 0 && x < 2 \\ &= \frac{1}{18}(2x+3) && 2 \le x \le 4 \\ &= 0 && x > 4 \end{aligned}$$

a probability density function? Find the probability that a variate having $f(x)$ as density function will fall in the interval $2 \le X \le 3$.

$\left[\textbf{Ans.: } \text{Yes, } \dfrac{4}{9}\right]$

10. A random variable X gives measurements x between 0 and 1 with a probability function

$$\begin{aligned} f(x) &= 12x^3 - 21x^2 + 10x && 0 \le x \le 1 \\ &= 0 && \text{otherwise} \end{aligned}$$

(i) Find $P\left(X \le \dfrac{1}{2}\right)$ and $P\left(X > \dfrac{1}{2}\right)$.

(ii) Find a number k such that $P(X \le k) = \dfrac{1}{2}$.

$\left[\textbf{Ans.: } \text{(i) } \dfrac{7}{16} \text{ (ii) } 0.452\right]$

11. The distribution function of a random variable X is given by

$$F(x) = \begin{cases} 1 - e^{-x^2} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Find the probability density function.

$$\left[\begin{aligned} \textbf{Ans.: } f(x) &= 2xe^{-x^2} && x > 0 \\ &= 0 && \text{otherwise} \end{aligned}\right]$$

12. The cdf of a continuous random variable X is given by

$$F(x)=\begin{cases}0 & x<0\\ x^2 & 0\le x\le 1\\ 1 & x>1\end{cases}$$

Find the pdf and $P\left(\frac{1}{2}\le X\le\frac{4}{5}\right)$.

[**Ans.:** 0.195]

13. Find the distribution function corresponding to the following probability density functions:

(i) $f(x)=\begin{cases}\frac{1}{2}x^2e^{-x} & 0\le x<\infty\\ 0 & \text{otherwise}\end{cases}$

(ii) $$\begin{aligned} f(x) &= x & & 0\le x\le 1\\ &= 2-x & & 1\le x\le 2\\ &= 0 & & \text{otherwise}\end{aligned}$$

(iii) $$\begin{aligned} f(x) &= \lambda(x-1)^4 & & 1\le x\le 3,\ \lambda>0\\ &= 0 & & \text{otherwise}\end{aligned}$$

$$\left[\begin{array}{l}\textbf{Ans.:}\ \text{(i)}\quad F(x)=\begin{cases}1-e^{-x}\left(1+x+\frac{x^2}{2}\right) & x\ge 0\\ 0 & \text{otherwise}\end{cases}\\ \qquad\text{(ii)}\quad F(x)=\begin{cases}0 & x<0\\ \frac{x^2}{2} & 0\le x\le 1\\ 2x-0.5x^2-1 & 1\le x\le 2\\ 1 & x>2\end{cases}\\ \qquad\text{(iii)}\quad \lambda=\frac{5}{32},\quad F(x)=\begin{cases}0 & x\le 1\\ \frac{5}{32}(x-1)^4 & 1\le x\le 3\\ 1 & x\ge 3\end{cases}\end{array}\right]$$

14. A continuous random variable X has the following probability density function

$$f(x)=\frac{a}{x^5}\qquad 2\le x\le 10$$

Determine the constant a, distribution function of X, and find the probability of the event $4\le x\le 7$.

$\left[\textbf{Ans.:}\ \frac{2500}{39},\ F(x)=\frac{625}{39}\left(\frac{1}{16}-\frac{1}{x^4}\right),\ 0.056\right]$

2.7 TWO-DIMENSIONAL DISCRETE RANDOM VARIABLES

In one-dimensional random variable, the outcome of any experiment had only one characteristic. In many situations, the outcome of a random experiment depends on two or more characteristics e.g., both voltage and current are measured in certain experiment.

Let X and Y be two random variables defined on the same sample space S, then the function (X, Y) that assigns a point in R^2 is called a two-dimensional random variable.

A two-dimensional random variable is said to be discrete if it takes at most a countable number of points in R^2. When (X, Y) is a two-dimensional discrete random variable, the possible values of (X, Y) may be represented as (x_i, y_j), $i = 1, 2, \ldots, m, \ldots;$ $j = 1, 2, \ldots, n, \ldots$

2.7.1 Joint Probability Mass Function

If (X, Y) is a two-dimensional discrete random variable, then the joint discrete function of X, Y, also called the joint probability mass function of X, Y, denoted by p_{XY} is defined by

$$p_{XY}(x_i, y_j)=P(X=x_i, Y=y_j)\quad \text{for a value of } (x_i, y_j) \text{ of } (X, Y)$$

and $$p_{XY}(x_i, y_j)=0,\quad \text{otherwise}$$

Following conditions should be satisfied for a function to be a probability mass function:

(i) $p_{XY}(x_i, y_j)\ge 0$, for all i and j

(ii) $\sum_{j=1}^{n}\sum_{i=1}^{m}p_{XY}(x_i, y_j)=1$.

2.7.2 Cumulative Distribution Function

If (X, Y) is a two-dimensional discrete random variable, then $F_{XY}(x, y) = P(X \le x, Y \le y)$ is called the cumulative distribution function (cdf) of (X, Y) and is defined by

$$F_{XY}(x,y)=\sum_{j=1}^{n}\sum_{i=1}^{m}p_{XY}(x_i, y_j)=\Sigma\Sigma p_{ij}$$

Properties of cdf

(i) $F(-\infty, y) = 0 = F(x,\infty)$ and $F(\infty,\infty) = 1$

(ii) $P(a < X < b,\ Y \le y) = F(b,y) - F(a,y)$

(iii) $P(X \le x\,,\ c < Y < d) = F(x,d) - F(x,c)$

(iv) $P(a < X < b,\ c < Y < d) = F(b,d) - F(a,d) - F(b,c) + F(a,c)$

(v) For a discrete random variable, $F_{XY}(x,y)$ will have step discontinuities. Derivatives at such discontinuities are not defined. At points of continuity,

$$\frac{\partial^2 F}{\partial x\,\partial y} = f(x,y)$$

2.7.3 Marginal Probability function

Let (X, Y) be a two-dimensional discrete random variable which takes up countable number of values (x_i, y_j). Then the probability distribution of X is given by

$$\begin{aligned} p_X(x_i) &= P(X = x_i) \\ &= P(X = x_i, Y = y_1) + P(X = x_i, Y = y_2) + \ldots + P(X = x_i, Y = y_m) \\ &= p_{i1} + p_{i2} + \cdots + p_{ij} + \cdots + p_{im} \\ &= \sum_{j=1}^{m} p_{ij} \\ &= \sum_{j=1}^{m} p(x_i, y_j) \\ &= p_{i*} \end{aligned}$$

and is known as marginal probability mass function or discrete marginal density function of X.

Similarly,

$$p_Y(y_j) = P(Y = y_j) = \sum_{i=1}^{n} p_{ij} = \sum_{i=1}^{n} p(x_i, y_j) = p_{*j}$$

is the marginal probability mass function of Y.

2.7.4 Conditional Probability Function

Let (X, Y) be a two-dimensional discrete random variable. Then the conditional discrete density function or conditional probability mass function of X, given $Y = y$, denoted by $p_{X/Y}(x/y)$ is defined as

$$p_{X/Y} = P(X = x / Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}, \text{ provided } P(Y = y) \neq 0.$$

The conditional probability mass function of Y, given $X = x$, denoted by $p_{Y/X}(y/x)$ is defined as:

$$p_{Y/X} = P(Y = y / X = x) = \frac{P(X = x, Y = y)}{P(X = x)}, \text{ provided } P(X = x) \neq 0.$$

A necessary and sufficient condition for the discrete random variables X and Y to be independent is

$$P(X = x_i, Y = y_j) = P(X = x_i)\, P(Y = y_j) \text{ for all values } (x_i, y_j) \text{ of } (X, Y)$$

Example 1

From the following table for bivariate distribution of (X, Y), find (i) $P(X \le 1)$ (ii) $P(Y \le 3)$ (iii) $P(X \le 1,\ Y \le 3)$, (iv) $P(X \le 1/Y \le 3)$ (v) $P(Y \le 3/X \le 1)$ (vi) $P(X + Y \le 4)$

X \ Y	1	2	3	4	5	6
0	0	0	$\frac{1}{32}$	$\frac{2}{32}$	$\frac{2}{32}$	$\frac{3}{32}$
1	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
2	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{64}$	0	$\frac{2}{64}$

Solution

Marginal distributions

X \ Y	1	2	3	4	5	6	$p_X(x)$
0	0	0	$\frac{1}{32}$	$\frac{2}{32}$	$\frac{2}{32}$	$\frac{3}{32}$	$\frac{8}{32}$
1	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{10}{16}$
2	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{64}$	0	$\frac{2}{64}$	$\frac{8}{64}$
$p_Y(y)$	$\frac{3}{32}$	$\frac{3}{32}$	$\frac{11}{64}$	$\frac{13}{64}$	$\frac{6}{32}$	$\frac{16}{64}$	$\Sigma p(x) = 1$ $\Sigma p(y) = 1$

(i) $$P(X \le 1) = P(X=0) + P(X=1)$$
$$= \frac{8}{32} + \frac{10}{16}$$
$$= \frac{7}{8}$$

(ii) $$P(Y \le 3) = P(Y=1) + P(Y=2) + P(Y=3)$$
$$= \frac{3}{32} + \frac{3}{32} + \frac{11}{64}$$
$$= \frac{23}{64}$$

(iii) $$P(X \le 1, Y \le 3) = P(X=0, Y=1) + P(X=0, Y=2) + P(X=0, Y=3) + P(X=1, Y=1) + P(X=1, Y=2) + P(X=1, Y=3)$$
$$= 0 + 0 + \frac{1}{32} + \frac{1}{16} + \frac{1}{16} + \frac{1}{8}$$
$$= \frac{9}{32}$$

(iv) $$P(X \le 1 / Y \le 3) = \frac{P(X \le 1, Y \le 3)}{P(Y \le 3)}$$
$$= \frac{\frac{9}{32}}{\frac{23}{64}}$$
$$= \frac{18}{23}$$

(v) $$P(Y \le 3 / X \le 1) = \frac{P(X \le 1, Y \le 3)}{P(X \le 1)}$$
$$= \frac{\frac{9}{32}}{\frac{7}{8}}$$
$$= \frac{9}{28}$$

(vi) $$P(X + Y \le 4) = P(X=0, Y=1) + P(X=0, Y=2) + P(X=0, Y=3) + P(X=0, Y=4) + P(X=1, Y=1) + P(X=1, Y=2) + P(X=1, Y=3) + P(X=2, Y=1) + P(X=2, Y=2)$$

$$= 0 + 0 + \frac{1}{32} + \frac{2}{32} + \frac{1}{16} + \frac{1}{16} + \frac{1}{8} + \frac{1}{32} + \frac{1}{32}$$
$$= \frac{13}{32}$$

Example 2

For the following joint distribution of X and Y, find the marginal distributions:

$Y \backslash X$	0	1	2
0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$
1	$\frac{3}{14}$	$\frac{3}{14}$	0
2	$\frac{1}{28}$	0	0

Solution

Marginal distributions

$Y \backslash X$	0	1	2	$p_Y(y)$
0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
1	$\frac{3}{14}$	$\frac{3}{14}$	0	$\frac{6}{14}$
2	$\frac{1}{28}$	0	0	$\frac{1}{28}$
$p_X(x)$	$\frac{10}{28}$	$\frac{15}{28}$	$\frac{3}{28}$	$\Sigma p(x) = 1$ $\Sigma p(y) = 1$

Marginal distributions of X

$$P(X=0) = P(X=0, Y=0) + P(X=0, Y=1) + P(X=0, Y=2)$$
$$= \frac{3}{28} + \frac{3}{14} + \frac{1}{28}$$
$$= \frac{10}{28}$$

$$P(X=1)=P(X=1,Y=0)+P(X=1,Y=1)+P(X=1,Y=2)$$
$$=\frac{9}{28}+\frac{3}{14}+0$$
$$=\frac{15}{28}$$

$$P(X=2)=P(X=2,Y=0)+P(X=2,Y=1)+P(X=2,Y=2)$$
$$=\frac{3}{28}+0+0$$
$$=\frac{3}{28}$$

Marginal distributions of Y

$$P(Y=0)=P(X=0,Y=0)+P(X=1,Y=0)+P(X=2,Y=0)$$
$$=\frac{3}{28}+\frac{9}{28}+\frac{3}{28}$$
$$=\frac{15}{28}$$

$$P(Y=1)=P(X=0,Y=1)+P(X=1,Y=1)+P(X=2,Y=1)$$
$$=\frac{3}{14}+\frac{3}{14}+0$$
$$=\frac{6}{14}$$

$$P(Y=2)=P(X=0,Y=2)+P(X=1,Y=2)+P(X=2,Y=2)$$
$$=\frac{1}{28}+0+0$$
$$=\frac{1}{28}$$

Example 3

The joint distribution of X and Y is given by

$$f(x,y)=\frac{x+y}{21},\quad x=1,2,3;\quad y=1,2$$

Find the marginal distributions.

Solution

Marginal distributions

Y \ X	1	2	3	$p_Y(y)$
1	$\frac{2}{21}$	$\frac{3}{21}$	$\frac{4}{21}$	$\frac{9}{21}$
2	$\frac{3}{21}$	$\frac{4}{21}$	$\frac{5}{21}$	$\frac{12}{21}$
$p_X(x)$	$\frac{5}{21}$	$\frac{7}{21}$	$\frac{9}{21}$	$\Sigma p(x)=1$ $\Sigma p(y)=1$

Marginal distributions of X

$$P(X=1)=P(X=1,Y=1)+P(X=1,Y=2)$$
$$=\frac{2}{21}+\frac{3}{21}$$
$$=\frac{5}{21}$$

$$P(X=2)=P(X=2,Y=1)+P(X=2,Y=2)$$
$$=\frac{3}{21}+\frac{4}{21}$$
$$=\frac{7}{21}$$

$$P(X=3)=P(X=3,Y=1)+P(X=3,Y=2)$$
$$=\frac{4}{21}+\frac{5}{21}$$
$$=\frac{9}{21}$$

Marginal distributions of Y

$$P(Y=1)=P(X=1,Y=1)+P(X=2,Y=1)+P(X=3,Y=1)$$
$$=\frac{2}{21}+\frac{3}{21}+\frac{4}{21}$$
$$=\frac{9}{21}$$

$$\begin{aligned}P(Y=2) &= P(X=1,Y=2)+P(X=2,Y=2)+P(X=3,Y=2)\\ &= \frac{3}{21}+\frac{4}{21}+\frac{5}{21}\\ &= \frac{12}{21}\end{aligned}$$

Example 4

Given is the joint distribution of X and Y

X Y	0	1	2
0	0.02	0.08	0.1
1	0.05	0.2	0.25
2	0.03	0.12	0.15

Find (i) marginal distributions (ii) the conditional distributions of X given $Y = 0$.

Solution

Marginal distributions

X Y	0	1	2	$p_Y(y)$
0	0.02	0.08	0.1	0.2
1	0.05	0.2	0.25	0.5
2	0.03	0.12	0.15	0.3
$p_X(x)$	0.1	0.4	0.5	$\Sigma p(x)=1$ $\Sigma p(y)=1$

Marginal distributions of X

$$\begin{aligned}P(X=0) &= P(X=0,Y=0)+P(X=0,Y=1)+P(X=0,Y=2)\\ &= 0.02+0.05+0.03\\ &= 0.1\end{aligned}$$

$$\begin{aligned}P(X=1) &= P(X=1,Y=0)+P(X=1,Y=1)+P(X=1,Y=2)\\ &= 0.08+0.2+0.12\\ &= 0.4\end{aligned}$$

$$\begin{aligned}P(X=2) &= P(X=2,Y=0)+P(X=2,Y=1)+P(X=2,Y=2)\\ &= 0.1+0.25+0.15\\ &= 0.5\end{aligned}$$

Marginal distributions of Y

$$\begin{aligned}P(Y=0) &= P(X=0,Y=0)+P(X=1,Y=0)+P(X=2,Y=0)\\ &= 0.02+0.08+0.1\\ &= 0.2\end{aligned}$$

$$\begin{aligned}P(Y=1) &= P(X=0,Y=1)+P(X=1,Y=1)+P(X=2,Y=1)\\ &= 0.05+0.2+0.25\\ &= 0.5\end{aligned}$$

$$\begin{aligned}P(Y=2) &= P(X=0,Y=2)+P(X=1,Y=2)+P(X=2,Y=2)\\ &= 0.03+0.12+0.15\\ &= 0.3\end{aligned}$$

Conditional distributions of X for $Y = 0$

$$P(X=0/Y=0)=\frac{P(X=0,Y=0)}{P(Y=0)}=\frac{0.02}{0.2}=0.1$$

$$P(X=1/Y=0)=\frac{P(X=1,Y=0)}{P(Y=0)}=\frac{0.08}{0.2}=0.4$$

$$P(X=2/Y=0)=\frac{P(X=2,Y=0)}{P(Y=0)}=\frac{0.1}{0.2}=0.5$$

$X=x$	0	1	2
$P(X=x/y=0)$	0.1	0.4	0.5

Example 5

The joint probability distribution of two random variables X and Y is given by

$$P(X=0,Y=1)=\frac{1}{3},\ P(X=1,Y=-1)=\frac{1}{3}\text{ and }P(X=1,Y=1)=\frac{1}{3}.$$

Find (i) marginal distributions of X and Y and (ii) the conditional probability distributions of X given $Y = 1$.

Solution

Marginal distributions

Y \ X	–1	0	1	*Marginal Y* $p_Y(y)$
–1	0	0	$\frac{1}{3}$	$\frac{1}{3}$
0	0	0	0	0
1	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$
Marginal X $p_X(x)$	0	$\frac{1}{3}$	$\frac{2}{3}$	$\Sigma p(y) = 1$ $\Sigma p(x) = 1$

Marginal distributions of X

$$\begin{aligned} P(X=-1) &= P(X=-1, Y=-1) + P(X=-1, Y=0) + P(X=-1, Y=1) \\ &= 0 \end{aligned}$$

$$\begin{aligned} P(X=0) &= P(X=0, Y=-1) + P(X=0, Y=0) + P(X=0, Y=1) \\ &= 0+0+\frac{1}{3} \\ &= \frac{1}{3} \end{aligned}$$

$$\begin{aligned} P(X=1) &= P(X=1, Y=-1) + P(X=1, Y=0) + P(X=1, Y=1) \\ &= \frac{1}{3}+0+\frac{1}{3} \\ &= \frac{2}{3} \end{aligned}$$

Marginal distributions of Y

$$\begin{aligned} P(Y=-1) &= P(X=-1, Y=-1) + P(X=0, Y=-1) + P(X=1, Y=-1) \\ &= 0+0+\frac{1}{3} \\ &= \frac{1}{3} \end{aligned}$$

$$\begin{aligned} P(Y=0) &= P(X=-1, Y=0) + P(X=0, Y=0) + P(X=1, Y=0) \\ &= 0 \end{aligned}$$

$$\begin{aligned} P(Y=1) &= P(X=-1, Y=1) + P(X=0, Y=1) + P(X=1, Y=1) \\ &= 0+\frac{1}{3}+\frac{1}{3} \\ &= \frac{2}{3} \end{aligned}$$

Conditional Probability distributions of X given Y = 1 is

$$P(X=x/Y=y) = \frac{P(X=x, Y=y)}{P(Y=y)}$$

$$P(X=-1/Y=1) = \frac{P(X=-1, Y=1)}{P(Y=1)} = 0$$

$$P(X=0/Y=1) = \frac{P(X=0, Y=1)}{P(Y=1)} = \frac{\frac{1}{3}}{\frac{2}{3}} = \frac{1}{2}$$

$$P(X=1/Y=1) = \frac{P(X=1, Y=1)}{P(Y=1)} = \frac{\frac{1}{3}}{\frac{2}{3}} = \frac{1}{2}$$

Example 6

If the joint probability mass function of (X, Y) is given by

$$P(x, y) = k(2x+3y), x = 0, 1, 2; y = 1, 2, 3$$

Find all the marginal probability distribution. Also, find the probability distribution of (X + Y).

Solution

$$P(x, y) = k(2x+3y)$$

Marginal distributions

Y \ X	0	1	2	$p_Y(y)$
1	3 k	5 k	7 k	15 k
2	6 k	8 k	10 k	24 k
3	9 k	11 k	13 k	33 k
$p_X(x)$	18 k	24 k	30 k	72 k

Solution

Marginal distributions

Y \ X	−1	0	1	Marginal Y $p_Y(y)$
−1	0	0	$\frac{1}{3}$	$\frac{1}{3}$
0	0	0	0	0
1	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$
Marginal X $p_X(x)$	0	$\frac{1}{3}$	$\frac{2}{3}$	$\Sigma p(y) = 1$ $\Sigma p(x) = 1$

Marginal distributions of X

$$\begin{aligned} P(X=-1) &= P(X=-1, Y=-1) + P(X=-1, Y=0) + P(X=-1, Y=1) \\ &= 0 \end{aligned}$$

$$\begin{aligned} P(X=0) &= P(X=0, Y=-1) + P(X=0, Y=0) + P(X=0, Y=1) \\ &= 0 + 0 + \frac{1}{3} \\ &= \frac{1}{3} \end{aligned}$$

$$\begin{aligned} P(X=1) &= P(X=1, Y=-1) + P(X=1, Y=0) + P(X=1, Y=1) \\ &= \frac{1}{3} + 0 + \frac{1}{3} \\ &= \frac{2}{3} \end{aligned}$$

Marginal distributions of Y

$$\begin{aligned} P(Y=-1) &= P(X=-1, Y=-1) + P(X=0, Y=-1) + P(X=1, Y=-1) \\ &= 0 + 0 + \frac{1}{3} \\ &= \frac{1}{3} \end{aligned}$$

$$\begin{aligned} P(Y=0) &= P(X=-1, Y=0) + P(X=0, Y=0) + P(X=1, Y=0) \\ &= 0 \end{aligned}$$

$$\begin{aligned} P(Y=1) &= P(X=-1, Y=1) + P(X=0, Y=1) + P(X=1, Y=1) \\ &= 0 + \frac{1}{3} + \frac{1}{3} \\ &= \frac{2}{3} \end{aligned}$$

Conditional Probability distributions of X given $Y = 1$ is

$$P(X = x / Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

$$P(X = -1 / Y = 1) = \frac{P(X = -1, Y = 1)}{P(Y = 1)} = 0$$

$$P(X = 0 / Y = 1) = \frac{P(X = 0, Y = 1)}{P(Y = 1)} = \frac{\frac{1}{3}}{\frac{2}{3}} = \frac{1}{2}$$

$$P(X = 1 / Y = 1) = \frac{P(X = 1, Y = 1)}{P(Y = 1)} = \frac{\frac{1}{3}}{\frac{2}{3}} = \frac{1}{2}$$

Example 6

If the joint probability mass function of (X, Y) is given by

$$P(x, y) = k(2x + 3y),\ x = 0, 1, 2;\ y = 1, 2, 3$$

Find all the marginal probability distribution. Also, find the probability distribution of $(X + Y)$.

Solution

$$P(x, y) = k(2x + 3y)$$

Marginal distributions

Y \ X	0	1	2	$p_Y(y)$
1	3 k	5 k	7 k	15 k
2	6 k	8 k	10 k	24 k
3	9 k	11 k	13 k	33 k
$p_X(x)$	18 k	24 k	30 k	72 k

$\because \quad \Sigma p(x) = \Sigma p(y) = 1$

$$72k = 1$$

$$k = \frac{1}{72}$$

Marginal distributions of X and Y

Y \ X	0	1	2	$p_Y(y)$
1	$\frac{3}{72}$	$\frac{5}{72}$	$\frac{7}{72}$	$\frac{15}{72}$
2	$\frac{6}{72}$	$\frac{8}{72}$	$\frac{10}{72}$	$\frac{24}{72}$
3	$\frac{9}{72}$	$\frac{11}{72}$	$\frac{13}{72}$	$\frac{33}{72}$
$p_X(x)$	$\frac{18}{72}$	$\frac{24}{72}$	$\frac{30}{72}$	1

Probability distribution of $(X + Y)$

$X + Y$	P
1	$p_{01} = \frac{3}{72}$
2	$p_{02} + p_{11} = \frac{11}{72}$
3	$p_{03} + p_{12} + p_{21} = \frac{24}{72}$
4	$p_{13} + p_{22} = \frac{21}{72}$
5	$p_{23} = \frac{13}{72}$
	Total = 1

Example 7

Let X and Y have the following marginal probability distributions:

X \ Y	0	1	2	$p_X(x)$
0	0.1	0.04	0.06	0.2
1	0.2	0.08	0.12	0.4
2	0.2	0.08	0.12	0.4
$p_Y(y)$	0.5	0.2	0.3	$\Sigma p(x) = 1$ $\Sigma p(y) = 1$

Solution

X and Y are independent, if $p_{ij} = p_{i*}\, p_{*j}$ for all i and j.

$$p_{0*} = 0.1 + 0.04 + 0.06 = 0.2$$
$$p_{1*} = 0.2 + 0.08 + 0.12 = 0.4$$
$$p_{2*} = 0.2 + 0.08 + 0.12 = 0.4$$
$$p_{*0} = 0.1 + 0.2 + 0.2 = 0.5$$
$$p_{*1} = 0.04 + 0.08 + 0.08 = 0.2$$
$$p_{*2} = 0.06 + 0.12 + 0.12 = 0.3$$

Now, $\quad p_{0*}\, p_{*0} = (0.2)\,(0.5) = 0.1 = p_{00}$

$$p_{0*}\, p_{*1} = (0.2)\,(0.2) = 0.04 = p_{01}$$
$$p_{0*}\, p_{*2} = (0.2)\,(0.3) = 0.06 = p_{02}$$

Similarly, it can be verified that

$$p_{1*}\, p_{*0} = p_{10};\ p_{1*}\, p_{*1} = p_{11};\ p_{1*} p_{*2} = p_{12}$$
$$p_{2*}\, p_{*0} = p_{20};\ p_{2*}\, p_{*1} = p_{21};\ p_{2*} p_{*2} = p_{22}$$

Hence, the random variables X and Y are independent.

EXERCISE 2.3

1. Find the marginal distributions of X and Y from the bivariate distribution of (X, Y) given below:

X \ Y	1	2
1	0.1	0.2
2	0.3	0.4

Ans.:

$X = x$	1	2
$P(X = x)$	0.3	0.7

$Y = y$	1	2
$P(Y = y)$	0.4	0.6

2. For the joint probability distribution of two random variables X and Y given below:

X \ Y	1	2	3	4	*Total*
1	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{10}{36}$
2	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{9}{36}$
3	$\frac{5}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{8}{36}$
4	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{5}{36}$	$\frac{9}{36}$
Total	$\frac{11}{36}$	$\frac{9}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	1

Find (i) marginal distributions of X and Y.

(ii) Conditional distributions of X given the value of $Y = 1$ and that of Y given the value of $X = 2$.

Ans.: (i)

Value of $X = x$	1	2	3	4
$P(X = x)$	$\frac{10}{36}$	$\frac{9}{36}$	$\frac{8}{36}$	$\frac{9}{36}$

Value of $Y = y$	1	2	3	4
$P(Y = y)$	$\frac{11}{36}$	$\frac{9}{36}$	$\frac{7}{36}$	$\frac{9}{36}$

(ii)

$Y = y$	1	2	3	4
$P(X = x/Y = 1)$	$\frac{4}{11}$	$\frac{1}{11}$	$\frac{5}{11}$	$\frac{1}{11}$

$Y = y$	1	2	3	4
$P(Y = y/X = 2)$	$\frac{1}{9}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{9}$

3. A two-dimensional random variable (X, Y) has the joint probability mass function $p(x, y) = \frac{2x + y}{27}$, where x and y can assume only the integer values 0, 1 and 2. Find the conditional distributions of Y for $X = x$.

Y \ X	0	1	2
0	0	$\frac{1}{3}$	$\frac{2}{3}$
1	$\frac{2}{9}$	$\frac{3}{9}$	$\frac{4}{9}$
2	$\frac{4}{15}$	$\frac{5}{15}$	$\frac{6}{15}$

4. Let X and Y have the following joint probability distribution:

Y \ X	0	1
0	0.1	0.15
1	0.2	0.3
2	0.1	0.15

Show that X and Y are independent.

5. The joint probability distribution of X and Y is given below:

Y \ X	1	2	3
2	$\frac{1}{8}$	$\frac{1}{24}$	$\frac{1}{12}$
4	$\frac{1}{4}$	$\frac{1}{4}$	0
6	$\frac{1}{8}$	$\frac{1}{24}$	$\frac{1}{12}$

Find $P(X < 4)$, $P(Y > 1)$, $P(X < 4/Y > 1)$, $P(2 \leq X \leq 5, Y > 1)$, $P(Y = 3/X = 2)$, $P(X + Y \leq 7)$.

$$\left[\textbf{Ans.:} \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}, \frac{3}{8}, \frac{1}{6}, \frac{19}{24}\right)\right]$$

6. For the following joint probability distribution of X and Y, find (i) marginal distributions of X and Y, (ii) conditional distributions of X given $Y = 2$, (iii) Are X and Y independent?

The conditional probability density function of Y, given $X = x$, denoted by $f(y/x)$ is defined as:

$$f(y/x) = \frac{f(x,y)}{f_X(x)}$$

A necessary and sufficient condition for the continuous random variables X and Y to be independent is

$$f(x,y) = f_X(x)\, f_Y(y)$$

Example 1

The joint probability density function of a two dimensional random variable is

$$\begin{aligned} f(x,y) &= \frac{1}{2}xe^{-y}, \quad 0<x<2, y>0 \\ &= 0, \qquad \text{otherwise} \end{aligned}$$

Find the cumulative distribution function.

Solution

The cumulative distribution function is given by

$$\begin{aligned} F(x,y) &= \int_{-\infty}^{y}\int_{-\infty}^{x} f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_0^y\int_0^x \frac{1}{2}xe^{-y}\mathrm{d}x\,\mathrm{d}y \\ &= \frac{1}{2}\int_0^y e^{-y}\left|\frac{x^2}{2}\right|_0^x \mathrm{d}y \\ &= \frac{1}{4}x^2\left|-e^{-y}\right|_0^y \\ &= \frac{1}{4}x^2(-e^{-y}+e^0) \\ &= \frac{1}{4}x^2(-e^{-y}+1) \end{aligned}$$

$$\begin{aligned} F(x,y) &= \frac{1}{4}x^2(1-e^{-y}), \quad 0<x<2, y>0 \\ &= 0 \qquad , \text{otherwise} \end{aligned}$$

Example 2

The joint probability density function of a two dimensional random variable (X, Y) is $f(x,y) = xe^{-x(y+1)}, x>0, y>0$. *Examine whether the variables X and Y are independent.*

Solution

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y \\ &= \int_0^{\infty} xe^{-x(y+1)}\mathrm{d}y \\ &= x\left|\frac{e^{-x(y+1)}}{-x}\right|_0^{\infty} \\ &= -(e^{-\infty}-e^{-x}) \\ &= e^{-x}, x>0 \end{aligned}$$

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x \\ &= \int_0^{\infty} xe^{-x(y+1)}\mathrm{d}x \\ &= \left|x\frac{e^{-x(y+1)}}{-(y+1)} - 1\frac{e^{-x(y+1)}}{(y+1)^2}\right|_0^{\infty} \\ &= \frac{1}{(y+1)^2}, y>0 \end{aligned}$$

$$f_X(x)\cdot f_Y(y) = e^{-x}\frac{1}{(y+1)^2}$$

$$f(x,y) = xe^{-x(y+1)}$$

$$f(x,y) \neq f_X(x)\cdot f_Y(y)$$

Hence, X and Y are not independent.

Example 3

Two random variables X and Y have the joint pdf

$$\begin{aligned} f(x,y) &= Ae^{-(2x+y)}, \quad x,y \geq 0 \\ &= 0 \qquad ,\text{otherwise} \end{aligned}$$

Find (i) *A* (ii) *marginal pdf of X and Y* (iii) $f(y/x)$

Solution

(i) Since $f(x, y)$ is a pdf,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\mathrm{d}x\,\mathrm{d}y = 1$$

$$\int_0^{\infty}\int_0^{\infty} Ae^{-(2x+y)}\mathrm{d}x\,\mathrm{d}y = 1$$

$$A\int_0^{\infty}\left[\int_0^{\infty} e^{-2x}\mathrm{d}x\right]e^{-y}\mathrm{d}y = 1$$

$$A\int_0^{\infty}\left|\frac{e^{-2x}}{-2}\right|_0^{\infty} e^{-y}\mathrm{d}y = 1$$

$$\frac{A}{-2}\int_0^{\infty}(e^{-\infty} - e^0)e^{-y}\mathrm{d}y = 1$$

$$\frac{A}{-2}\int_0^{\infty}(-1)e^{-y}\mathrm{d}y = 1$$

$$\frac{A}{2}\left|-e^{-y}\right|_0^{\infty} = 1$$

$$\frac{A}{2}(-e^{-\infty} + e^0) = 1$$

$$A = 2$$

(ii) $$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\mathrm{d}y$$

$$= \int_0^{\infty} Ae^{-(2x+y)}\mathrm{d}y$$

$$= 2\left|-e^{-(2x+y)}\right|_0^{\infty}$$

$$= 2(-e^{-\infty} + e^{-2x})$$

$$= 2e^{-2x}, x \geq 0$$

$$\therefore \quad f_X(x) = 2e^{-2x}, \quad x \geq 0$$

$$= 0 \quad , \quad x < 0$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x$$

$$= \int_0^{\infty} Ae^{-(2x+y)}\mathrm{d}x$$

$$= 2\left|\frac{e^{-(2x+y)}}{-2}\right|_0^{\infty}$$

$$= -\frac{2}{2}(e^{-\infty} - e^{-y})$$

$$= -(0 - e^{-y})$$

$$= e^{-y}, y \geq 0$$

$$\therefore f_Y(y) = e^{-y}, \ y \geq 0$$

$$= 0 \ , \ y < 0$$

(iii) $$f(y/x) = \frac{f(x,y)}{f_X(x)}$$

$$= \frac{2e^{-(2x+y)}}{2e^{-2x}}$$

$$= e^{-y}, \quad y \geq 0$$

Example 4

The joint probability distribution of X and Y is given by

$$f(x,y) = \frac{6-x-y}{8}, \quad 0 < x < 2, \quad 2 < y < 4$$

$$= 0, \qquad \text{otherwise}$$

Find $f(y/x = 2)$.

Solution

$$f(y/x) = \frac{f(x,y)}{f_X(x)}$$

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y$$

$$= \int_2^4 \frac{6-x-y}{8}\,\mathrm{d}y$$

$$= \frac{1}{8}\left|6y - xy - \frac{y^2}{2}\right|_2^4$$

$$= \frac{1}{8}[(24-4x-8)-(12-2x-2)]$$

$$= \frac{1}{8}(6-2x)$$

$$f(y/x) = \frac{6-x-y}{6-2x}$$

Putting $x = 2$,

$$f(y/x=2) = \frac{4-y}{2}$$

Example 5

The joint pdf of a two dimensional variable (X, Y) is given by

$$f(x,y) = kxye^{-(x^2+y^2)}, \ x>0, y>0.$$

Find the value of k and prove that X and Y are independent.

Solution

Since $f(x, y)$ is a pdf,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = 1$$

$$\int_0^{\infty}\int_0^{\infty} k\,x\,y\,e^{-(x^2+y^2)}\mathrm{d}x\,\mathrm{d}y = 1$$

$$k\int_0^{\infty} ye^{-y^2}\mathrm{d}y \cdot \int_0^{\infty} xe^{-x^2}\mathrm{d}x = 1 \qquad \text{...(1)}$$

Putting $x^2 = t, x = \sqrt{t}, \mathrm{d}x = \frac{1}{2\sqrt{t}}\mathrm{d}t$

When $x = 0,\ t = 0$

When $x = \infty$, $t = \infty$

$$\int_0^{\infty} xe^{-x^2}\mathrm{d}x = \int_0^{\infty}\sqrt{t}e^{-t}\frac{1}{2\sqrt{t}}\mathrm{d}t$$

$$= \frac{1}{2}\left|-e^{-t}\right|_0^{\infty}$$

$$= \frac{1}{2}(-e^{-\infty}+e^0)$$

$$= \frac{1}{2}$$

Similarly, $\int_0^{\infty} ye^{-y^2}\mathrm{d}y = \frac{1}{2}$

Putting both integral values in Eq. (1),

$$k \cdot \frac{1}{2} \cdot \frac{1}{2} = 1$$

$$k = 4$$

If X and Y are independent,

$$f_X(x) \cdot f_Y(y) = f(x,y)$$

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y$$

$$= \int_0^{\infty} k\,x\,y\,e^{-(x^2+y^2)}\mathrm{d}y$$

$$= k\,x\,e^{-x^2}\int_0^{\infty} y\,e^{-y^2}\mathrm{d}y$$

$$= 4x\,e^{-x^2} \cdot \frac{1}{2}$$

$$= 2x\,e^{-x^2}, x>0$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x$$

$$= \int_0^{\infty} k\,x\,y\,e^{-(x^2+y^2)}\mathrm{d}x$$

$$= k\,y\,e^{-y^2}\int_0^{\infty} xe^{-x^2}\mathrm{d}x$$

$$= 4y\,e^{-y^2} \cdot \frac{1}{2}$$

$$= 2ye^{-y^2}, y>0$$

$$f_X(x) \cdot f_Y(y) = 2x\,e^{-x^2} \cdot 2y\,e^{-y^2}, x, y>0$$

$$= 4xy\,e^{-(x^2+y^2)}$$

$$= f(x,y), \quad x>0, y>0$$

Hence, X and Y are independent.

Example 6

The joint probability density function of a two dimensional random variable (X, Y) is

$$f(x,y) = kx(x-y), \quad 0<x<2, -x<y<x$$
$$= 0 \quad , \quad \text{elsewhere}$$

Find (i) k (ii) $f_X(x)$ (iii) $f_Y(y)$ (iv) $f(y/x)$

Solution

Since $f(x, y)$ is a probability density function,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = 1$$

$$\int_0^2\int_{-x}^{x} kx(x-y)\,\mathrm{d}y\,\mathrm{d}x = 1$$

$$k\int_0^2 x\left|xy-\frac{y^2}{2}\right|_{-x}^{x}\mathrm{d}x = 1$$

$$k\int_0^2 x\left[\left(x^2-\frac{x^2}{2}\right)-\left(-x^2-\frac{x^2}{2}\right)\right]\mathrm{d}x = 1$$

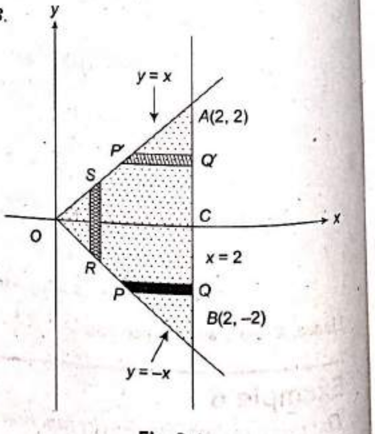
$$k\int_0^2 2x^3\,\mathrm{d}x = 1$$

$$k\left|\frac{2x^4}{4}\right|_0^2 = 1$$

$$k(8) = 1$$

$$k = \frac{1}{8}$$

(ii) The region of integration is ΔOAB.
In ΔOAB, along vertical strip RS,
Limits of y : $y = -x$ to $y = x$
and x varies from $x = 0$ to $x = 2$.



$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y$$
$$= \int_{-x}^{x} kx(x-y)\,\mathrm{d}y$$
$$= kx\left|xy-\frac{y^2}{2}\right|_{-x}^{x}$$
$$= kx\left[\left(x^2-\frac{x^2}{2}\right)-\left(-x^2-\frac{x^2}{2}\right)\right]$$
$$= kx\,(2x^2)$$
$$= \frac{1}{8}(2x^3)$$

$$= \frac{x^3}{4}, \quad 0<x<2$$

(iii) For limits of x, ΔOAB is divided into two parts, ΔOBC and ΔOAC.
In ΔOBC, along horizontal strip PQ,
Limits of x: $x = -y$ to $x = 2$ and y varies from $y = -2$ to $y = 0$.
In ΔOAC, along horizontal strip $P'Q'$,
Limits of x: $x = y$ to $x = 2$ and y varies from $y = 0$ to $y = 2$.

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x$$
$$= \int_{-y}^{2} kx(x-y)\,\mathrm{d}x, \quad -2 \le y \le 0$$
$$= \int_{y}^{2} kx(x-y)\,\mathrm{d}x, \quad 0 \le y \le 2$$

Now,

$$\int_{-y}^{2} kx(x-y)\,\mathrm{d}x = k\left|\frac{x^3}{3}-\frac{x^2y}{2}\right|_{-y}^{2}$$
$$= k\left[\left(\frac{8}{3}-2y\right)-\left(-\frac{y^3}{3}-\frac{y^3}{2}\right)\right]$$
$$= \frac{1}{8}\left(\frac{8}{3}-2y+\frac{5y^3}{6}\right)$$
$$= \frac{1}{3}-\frac{y}{4}+\frac{5}{48}y^3$$

Also,

$$\int_{y}^{2} kx(x-y)\,\mathrm{d}x = k\left|\frac{x^3}{3}-\frac{x^2y}{2}\right|_{y}^{2}$$
$$= k\left[\left(\frac{8}{3}-2y\right)-\left(\frac{y^3}{3}-\frac{y^3}{2}\right)\right]$$
$$= \frac{1}{8}\left(\frac{8}{3}-2y+\frac{y^3}{6}\right)$$
$$= \frac{1}{3}-\frac{y}{4}+\frac{y^3}{48}$$

Hence, $f_Y(y) = \frac{1}{3}-\frac{y}{4}+\frac{5}{48}y^3, \quad -2 \le y \le 0$

$$= \frac{1}{3}-\frac{y}{4}+\frac{y^3}{48}, \quad 0 \le y \le 2$$

(iv) $f(y/x)=\dfrac{f(x,y)}{f_X(x)}$

$$=\frac{\frac{1}{8}x(x-y)}{\frac{x^3}{4}}$$

$$=\frac{x-y}{2x^2},\quad -x<y<x$$

Example 7

The joint pdf of a two-dimensional random variable (X, Y) is given by

$$f_{XY}(x,y)=\frac{8}{9}xy,\qquad 1\le y\le 2,\ \ 1\le x\le y$$
$$=0\ ,\qquad \text{otherwise}$$

Find the marginal density function of X and Y.

Solution

The region of integration is ΔABC.
In ΔABC, along vertical strip PQ, limits of y: $y=x$ to $y=2$ and x varies from $x=1$ to $x=2$.
Marginal density function of X is

$$f_X(x)=\int_{-\infty}^{\infty}f(x,y)\mathrm{d}y$$
$$=\int_x^2\frac{8}{9}xy\ \mathrm{d}y$$
$$=\frac{8}{9}x\left|\frac{y^2}{2}\right|_x^2$$
$$=\frac{4}{9}x(4-x^2),\qquad 1\le x\le 2$$

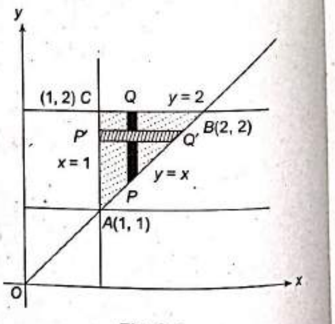


Fig. 2.2

In ΔABC, along horizontal strip $P'Q'$, limits of x : $x=1$ to $x=y$ and y varies from $y=1$ to $y=2$.

Marginal density function of Y is

$$f_Y(y)=\int_{-\infty}^{\infty}f(x,y)\mathrm{d}x$$
$$=\int_1^y\frac{8}{9}xy\ \mathrm{d}x$$

$$=\frac{8}{9}y\left|\frac{x^2}{2}\right|_1^y$$
$$=\frac{4}{9}y(y^2-1),\qquad 1\le y\le 2$$

Example 8

If the joint distribution function of X and Y is given by

$$F(x,y)=(1-e^{-x})(1-e^{-y}),\qquad x>0,y>0$$
$$=0\qquad ,\qquad \text{otherwise}$$

Find $f_X(x), f_Y(y)$ (ii) *Are X and Y are independent* (iii) *Find* $P(1<X<3,\ 1<Y<2)$.

Solution

$$F(x,y)=(1-e^{-x})(1-e^{-y})$$

The joint pdf is given by

$$f(x,y)=\frac{\partial^2 F}{2x\partial y}$$
$$=\frac{\partial}{\partial x}\left(\frac{\partial F}{\partial y}\right)$$
$$=\frac{\partial}{\partial x}\left[\frac{\partial}{\partial y}(1-e^{-x})(1-e^{-y})\right]$$
$$=\frac{\partial}{\partial x}(1-e^{-x})(e^{-y})$$
$$=e^{-x}e^{-y}$$
$$=e^{-(x+y)},x>0,y>0$$

$$\therefore\qquad f(x,y)=e^{-(x+y)},\quad x>0,y>0$$
$$=0,\qquad ,\ \text{otherwise}$$

(i) $f_X(x)=\int_{-\infty}^{\infty}f(x,y)\ \mathrm{d}y$

$$=\int_0^{\infty}e^{-(x+y)}\mathrm{d}y$$
$$=\left|-e^{-(x+y)}\right|_0^{\infty}$$
$$=(-e^{-\infty}+e^{-x})$$
$$=e^{-x},x>0$$

$$
\begin{aligned}
f_Y(y) &= \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x \\
&= \int_0^{\infty} e^{-(x+y)}\mathrm{d}x \\
&= \left|-e^{-(x+y)}\right|_0^{\infty} \\
&= (-e^{-\infty} + e^{-y}) \\
&= e^{-y}, x > 0
\end{aligned}
$$

(ii) $f_X(x) \cdot f_Y(y) = e^{-x} \cdot e^{-y}$

$$
\begin{aligned}
&= e^{-(x+y)}, x > 0, y > 0 \\
&= f(x,y)
\end{aligned}
$$

Hence, X and Y are independent.

(iii) Since X and Y are independent,

$$
\begin{aligned}
P(1 < X < 3, 1 < Y < 2) &= P(1 < X < 3) \cdot P(1 < Y < 2) \\
&= \int_1^3 f_X(x)\,\mathrm{d}x \cdot \int_1^2 f_Y(y)\,\mathrm{d}y \\
&= \int_1^3 e^{-x}\,\mathrm{d}x \cdot \int_1^2 e^{-y}\,\mathrm{d}y \\
&= \left|-e^{-x}\right|_1^3 \cdot \left|-e^{-y}\right|_1^2 \\
&= (-e^{-3} + e^{-1}) \cdot (-e^{-2} + e^{-1}) \\
&= e^{-5} - e^{-4} - e^{-3} - e^{-2}
\end{aligned}
$$

Example 9

The joint probability density of two random variables is given by

$$
\begin{aligned}
f(x,y) &= 15e^{-3x-5y}, \quad x > 0, y > 0 \\
&= 0 \quad\quad , \text{ elsewhere}
\end{aligned}
$$

Find (i) $P(1 < X < 2, 0.2 < Y < 0.3)$ (ii) $P(X < 2, Y > 0.2)$ (iii) *marginal probability density functions of X and Y.*

Solution

(i)
$$
\begin{aligned}
P(1 < X < 2, 0.2 < Y < 0.3) &= \int_{0.2}^{0.3}\int_1^2 f(x,y)\,\mathrm{d}x\mathrm{d}y \\
&= \int_{0.2}^{0.3}\int_1^2 15e^{-3x-5y}\,\mathrm{d}x\mathrm{d}y \\
&= 15\int_{0.2}^{0.3} e^{-5y}\left[\int_1^2 e^{-3x}\mathrm{d}x\right]\mathrm{d}y
\end{aligned}
$$

$$
\begin{aligned}
&= 15\int_{0.2}^{0.3} e^{-5y}\left|\frac{e^{-3x}}{-3}\right|_1^2 \mathrm{d}y \\
&= -5\int_{0.2}^{0.3} e^{-5y}(e^{-6} - e^{-3})\mathrm{d}y \\
&= -5(e^{-6} - e^{-3})\left|\frac{e^{-5y}}{-5}\right|_{0.2}^{0.3} \\
&= (e^{-6} - e^{-3})(e^{-1.5} - e^{-1.0}) \\
&= 6.84 \times 10^{-3}
\end{aligned}
$$

(ii)
$$
\begin{aligned}
P(X < 2, Y > 0.2) &= \int_{0.2}^{\infty}\int_0^2 f(x,y)\,\mathrm{d}x\mathrm{d}y \\
&= \int_{0.2}^{\infty}\int_0^2 15e^{-3x-5y}\,\mathrm{d}x\mathrm{d}y \\
&= 15\int_{0.2}^{\infty}\left[\int_0^2 e^{-3x}\mathrm{d}x\right]e^{-5y}\mathrm{d}y \\
&= 15\int_{0.2}^{\infty}\left|\frac{e^{-3x}}{-3}\right|_0^2 e^{-5y}\mathrm{d}y \\
&= -5\int_{0.2}^{\infty}(e^{-6} - 1)e^{-5y}\mathrm{d}y \\
&= -5(e^{-6} - 1)\left|\frac{e^{-5y}}{-5}\right|_{0.2}^{\infty} \\
&= (e^{-6} - 1)(e^{-\infty} - e^{-1.0}) \\
&= (e^{-6} - 1)(-e^{-1.0}) \\
&= 0.367
\end{aligned}
$$

(iii) The region of integration is the first quadrant. Hence, x and y both varies from 0 to ∞.

$$
\begin{aligned}
f_X(x) &= \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y \\
&= \int_0^{\infty} 15e^{-3x-5y}\mathrm{d}y \\
&= 15e^{-3x}\left|\frac{e^{-5y}}{-5}\right|_0^{\infty}
\end{aligned}
$$

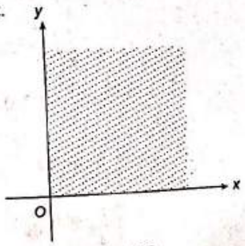


Fig. 2.3

$$= -3e^{-3x}(e^{-\infty} - e^{0})$$

$$= 3e^{-3x}, \quad x > 0$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x$$

$$= \int_{0}^{\infty} 15e^{-3x-5y}\,\mathrm{d}y$$

$$= 15e^{-5y}\left|\frac{e^{-3x}}{-3}\right|_{0}^{\infty}$$

$$= -5e^{-5y}(e^{-\infty} - e^{0})$$

$$= 5e^{-5y}, \quad y > 0$$

Example 10

The joint pdf of (X, Y) is given by

$$f(x,y) = \frac{1}{4}e^{-|x|-|y|}, \quad -\infty < x < \infty, -\infty < y < \infty$$

(i) Are X and Y independent?

(ii) Find the probability that $X \le 1$ and $Y < 0$.

Solution

$$|x| = -x, \quad -\infty < x \le 0$$

$$= x, \quad 0 \le x < \infty$$

Similarly, $|y| = -y, \quad -\infty < y \le 0$

$$= y, \quad 0 \le y < \infty$$

$$\therefore \quad f(x,y) = \frac{1}{4}e^{-|x|-|y|}$$

$$= \frac{1}{4}e^{x+y}, -\infty < x \le 0, -\infty < y \le 0$$

$$= \frac{1}{4}e^{-x-y}, 0 \le x < 0, 0 \le y < \infty$$

(i) $f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y$

$$= \int_{-\infty}^{\infty} \frac{1}{4}e^{-|x|-|y|}\,\mathrm{d}y$$

$$= \frac{1}{4}e^{-|x|}\int_{-\infty}^{\infty} e^{-|y|}\mathrm{d}y$$

$$= \frac{1}{4}e^{-|x|}\left[\int_{-\infty}^{0} e^{y}\mathrm{d}y + \int_{0}^{\infty} e^{-y}\mathrm{d}y\right]$$

$$= \frac{1}{4}e^{-|x|}\left[\left|e^{y}\right|_{-\infty}^{0} + \left|-e^{-y}\right|_{0}^{\infty}\right]$$

$$= \frac{1}{4}e^{-|x|}(1+1)$$

$$= \frac{1}{2}e^{-|x|}, \quad -\infty < x < \infty$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x$$

$$= \int_{-\infty}^{\infty} \frac{1}{4}e^{-|x|-|y|}\,\mathrm{d}x$$

$$= \frac{1}{4}e^{-|y|}\int_{-\infty}^{\infty} e^{-|x|}\mathrm{d}x$$

$$= \frac{1}{4}e^{-|y|}\left[\int_{-\infty}^{0} e^{x}\mathrm{d}x + \int_{0}^{\infty} e^{-x}\mathrm{d}x\right]$$

$$= \frac{1}{4}e^{-|y|}\left[\left|e^{x}\right|_{-\infty}^{0} + \left|-e^{-x}\right|_{0}^{\infty}\right]$$

$$= \frac{1}{4}e^{-|y|}(1+1)$$

$$= \frac{1}{2}e^{-|y|}, \quad -\infty < y < \infty$$

$$f_X(x) \cdot f_Y(y) = \frac{1}{2}e^{-|x|} \cdot \frac{1}{2}e^{-|y|}$$

$$= \frac{1}{4}e^{-|x|-|y|}, -\infty < x < \infty, -\infty < y < \infty$$

$$= f(x,y)$$

Hence, X and Y are independent.

(ii) $P(X \le 1, Y < 0) = \int_{-\infty}^{0}\int_{-\infty}^{1} f(x,y)\,\mathrm{d}x\,\mathrm{d}y$

$$= \int_{-\infty}^{0}\int_{-\infty}^{1} \frac{1}{4}e^{-|x|-|y|}\,\mathrm{d}x\,\mathrm{d}y$$

$$=\frac{1}{4}\int_{-\infty}^{0} e^{-|y|}\left[\int_{-\infty}^{0} e^{x}\mathrm{d}x+\int_{0}^{1} e^{-x}\mathrm{d}x\right]\mathrm{d}y$$

$$=\frac{1}{4}\int_{-\infty}^{0} e^{-|y|}\left[\left|e^{x}\right|_{-\infty}^{0}+\left|-e^{-x}\right|_{0}^{1}\right]\mathrm{d}y$$

$$=\frac{1}{4}\int_{-\infty}^{0} e^{-|y|}(1-e^{-1}+1)\,\mathrm{d}y$$

$$=\frac{1}{4}(2-e^{-1})\int_{-\infty}^{0} e^{y}\mathrm{d}y$$

$$=\frac{1}{4}(2-e^{-1})\left|e^{y}\right|_{-\infty}^{0}$$

$$=\frac{1}{4}(2-e^{-1})(1)$$

$$=\frac{1}{4}(2-e^{-1})$$

Example 11

The joint pdf of (X, Y) is given by

$$f(x,y)=ke^{-x}\cos y,\quad 0\le x\le 2, 0\le y\le\frac{\pi}{2}$$
$$=0\qquad,\qquad \text{otherwise}$$

Find (i) k (ii) $P\left(X+Y\ge\frac{\pi}{2}\right)$.

Solution

(i) Since $f(x, y)$ is a pdf,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x\,\mathrm{d}y=1$$

$$\int_{0}^{\frac{\pi}{2}}\int_{0}^{2} k\,e^{-x}\cos y\,\mathrm{d}x\,\mathrm{d}y=1$$

$$k\int_{0}^{\frac{\pi}{2}}\cos y\left|-e^{-x}\right|_{0}^{2}\mathrm{d}y=1$$

$$k\int_{0}^{\frac{\pi}{2}}\cos y\,(-e^{-2}+1)\,\mathrm{d}y=1$$

$$k(1-e^{-2})\left|\sin y\right|_{0}^{\frac{\pi}{2}}=1$$

$$k(1-e^{-2})(1)=1$$

$$k=\frac{1}{1-e^{-2}}$$

(ii) $P\left(X+Y\ge\frac{\pi}{2}\right)=1-P\left(X+Y<\frac{\pi}{2}\right)$

The region of integration $x + y < 1$ is the ΔOAB. In ΔOAB, along horizontal strip $P'Q'$,

Limits of $x: x=0$ to $x=\frac{\pi}{2}-y$

Limits of $y: y=0$ to $y=\frac{\pi}{2}$

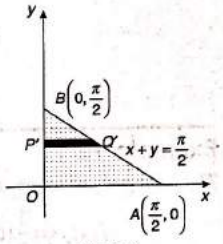

y
B$\left(0,\frac{\pi}{2}\right)$
P′
Q′
$x+y=\frac{\pi}{2}$
O
x
A$\left(\frac{\pi}{2},0\right)$

Fig. 2.4

$$P(X+Y<1)=\int_{0}^{\frac{\pi}{2}}\int_{0}^{\frac{\pi}{2}-y} ke^{-x}\cos y\,\mathrm{d}x\,\mathrm{d}y$$

$$=k\int_{0}^{\frac{\pi}{2}}\cos y\left|-e^{-x}\right|_{0}^{\frac{\pi}{2}-y}\mathrm{d}y$$

$$=k\int_{0}^{\frac{\pi}{2}}\cos y\left[-e^{-\left(\frac{\pi}{2}-y\right)}+e^{0}\right]\mathrm{d}y$$

$$=k\int_{0}^{\frac{\pi}{2}}\cos y\left[-e^{-\frac{\pi}{2}}e^{y}+1\right]\mathrm{d}y$$

$$=k\left[-e^{-\frac{\pi}{2}}\int_{0}^{\frac{\pi}{2}} e^{y}\cos y\,\mathrm{d}y+\int_{0}^{\frac{\pi}{2}}\cos y\,\mathrm{d}y\right]$$

$$=k\left[-e^{-\frac{\pi}{2}}\left|\frac{e^{y}}{1+1}(\cos y+\sin y)\right|_{0}^{\frac{\pi}{2}}+\left|\sin y\right|_{0}^{\frac{\pi}{2}}\right]$$

$$=k\left[-e^{-\frac{\pi}{2}}\left\{\frac{e^{\frac{\pi}{2}}}{2}\left(\cos\frac{\pi}{2}+\sin\frac{\pi}{2}\right)-\frac{1}{2}\right\}+\sin\frac{\pi}{2}\right.$$

$$=k\left[-e^{-\frac{\pi}{2}}\left\{\frac{e^{\frac{\pi}{2}}}{2}(1)-\frac{1}{2}\right\}+1\right]$$

$$=k\left(-\frac{1}{2}+\frac{e^{-\frac{\pi}{2}}}{2}+1\right)$$

$$= \frac{k}{2}\left(1+e^{-\frac{\pi}{2}}\right)$$

$$P\left(X+Y \geq \frac{\pi}{2}\right) = 1-\frac{k}{2}\left(1+e^{-\frac{\pi}{2}}\right)$$

$$= 1-\frac{\left(1+e^{-\frac{\pi}{2}}\right)}{2(1-e^{-2})}$$

Example 12

The joint p.d.f of a two-dimensional random variable (X, Y) is given by

$$f(x, y) = \frac{1}{8}(6-x-y), \qquad 0<x<2, \quad 2<y<4$$
$$= 0 \qquad , \qquad \text{otherwise}$$

Find (i) $P(X<1, Y<3)$ (ii) $P(X<1/Y<3)$.

Solution

(i) $P(X<1, Y<3) = \int_2^3\int_0^1 f(x,y)\,dx\,dy$

$$= \int_2^3\int_0^1 \frac{1}{8}(6-x-y)\,dx\,dy$$

$$= \frac{1}{8}\int_2^3 \left|6x-\frac{x^2}{2}-xy\right|_0^1 dy$$

$$= \frac{1}{8}\int_2^3 \left(6-\frac{1}{2}-y\right)dy$$

$$= \frac{1}{8}\int_2^3 \left(\frac{11}{2}-y\right)dy$$

$$= \frac{1}{8}\left|\frac{11}{2}y-\frac{y^2}{2}\right|_2^3$$

$$= \frac{1}{8}\left[\left(\frac{33}{2}-\frac{9}{2}\right)-(11-2)\right]$$

$$= \frac{3}{8}$$

(ii) $P(X<1/Y<3) = \frac{P(X<1, Y<3)}{P(Y<3)}$ ...(1)

$$P(Y<3) = \int_2^3\int_0^2 f(x,y)\,dx\,dy$$

$$= \int_2^3\int_0^2 \frac{1}{8}(6-x-y)\,dx\,dy$$

$$= \frac{1}{8}\int_2^3 \left|6x-\frac{x^2}{2}-xy\right|_0^2 dy$$

$$= \frac{1}{8}\int_2^3 (12-2-2y)\,dy$$

$$= \frac{1}{8}\int_2^3 (10-2y)\,dy$$

$$= \frac{1}{8}\left|10y-y^2\right|_2^3$$

$$= \frac{1}{8}\left[(30-9)-(20-4)\right]$$

$$= \frac{5}{8}$$

Substituting in Eq (1),

$$P(X<1/Y<3) = \frac{\left(\frac{3}{8}\right)}{\left(\frac{5}{8}\right)} = \frac{3}{5}$$

Example 13

The joint pdf of a two-dimensional random variable (X, Y) is given by

$$f_{XY}(x,y) = \begin{cases} xy^2+\frac{x^2}{8}, & 0<x<2, \quad 0<y<1 \\ 0 \quad , & \text{Otherwise} \end{cases}$$

Find (i) $P(X>1)$ (ii) $P\left(Y<\frac{1}{2}\right)$ (iii) $P\left(X>1/Y<\frac{1}{2}\right)$

(iv) $P\left(Y<\frac{1}{2}/X>1\right)$ (v) $P(X<Y)$ (vi) $P(X+Y \leq 1)$

Solution

(i) $$P(X>1)=\int_0^1\int_1^2 f(x,y)\,dx\,dy$$

$$=\int_0^1\int_1^2\left(xy^2+\frac{x^2}{8}\right)dx\,dy$$

$$=\int_0^1\left|\frac{x^2y^2}{2}+\frac{x^3}{24}\right|_1^2 dy$$

$$=\int_0^1\left(2y^2+\frac{1}{3}\right)-\left(\frac{y^2}{2}+\frac{1}{24}\right)dy$$

$$=\int_0^1\left(\frac{3y^2}{2}+\frac{7}{24}\right)dy$$

$$=\left|\frac{y^3}{2}+\frac{7y}{24}\right|_0^1$$

$$=\frac{1}{2}+\frac{7}{24}$$

$$=\frac{19}{24}$$

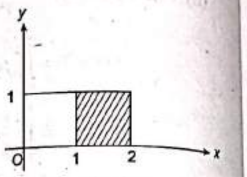


Fig. 2.5

(ii) $$P\left(Y<\frac{1}{2}\right)=\int_0^{\frac{1}{2}}\int_0^2 f(x,y)\,dx\,dy$$

$$=\int_0^{\frac{1}{2}}\int_0^2\left(xy^2+\frac{x^2}{8}\right)dx\,dy$$

$$=\int_0^{\frac{1}{2}}\left|\frac{x^2y^2}{2}+\frac{x^3}{24}\right|_0^2 dy$$

$$=\int_0^{\frac{1}{2}}\left(2y^2+\frac{1}{3}\right)dy$$

$$=\left|\frac{2y^3}{3}+\frac{1}{3}y\right|_0^{\frac{1}{2}}$$

$$=\frac{1}{12}+\frac{1}{6}$$

$$=\frac{1}{4}$$

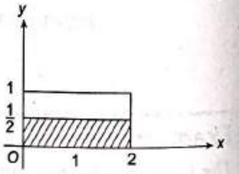


Fig. 2.6

(iii) $$P\left(X>1,Y<\frac{1}{2}\right)=\int_0^{\frac{1}{2}}\int_1^2 f(x,y)\,dx\,dy$$

$$=\int_0^{\frac{1}{2}}\int_1^2\left(xy^2+\frac{x^2}{8}\right)dx\,dy$$

$$=\int_0^{\frac{1}{2}}\left|\frac{x^2y^2}{2}+\frac{x^3}{24}\right|_1^2 dy$$

$$=\int_0^{\frac{1}{2}}\left(2y^2+\frac{1}{3}\right)-\left(\frac{y^2}{2}+\frac{1}{24}\right)dy$$

$$=\int_0^{\frac{1}{2}}\left(\frac{3y^2}{2}+\frac{7}{24}\right)dy$$

$$=\left|\frac{y^3}{2}+\frac{7y}{24}\right|_0^{\frac{1}{2}}$$

$$=\frac{1}{16}+\frac{7}{48}$$

$$=\frac{5}{24}$$

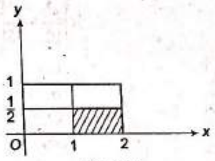


Fig. 2.7

$$P\left(X>1/Y<\frac{1}{2}\right)=\frac{P\left(X>1,Y<\frac{1}{2}\right)}{P\left(Y<\frac{1}{2}\right)}=\frac{\frac{5}{24}}{\frac{1}{4}}=\frac{5}{6}$$

(iv) $$P\left(Y<\frac{1}{2}/X>1\right)=\frac{P\left(X>1,Y<\frac{1}{2}\right)}{P(X>1)}=\frac{\frac{5}{24}}{\frac{19}{24}}=\frac{5}{19}$$

(v) $$P(X<Y)=\int_0^1\int_0^y f(x,y)\,dx\,dy$$

$$=\int_0^1\int_0^y\left(xy^2+\frac{x^2}{8}\right)dx\,dy$$

$$=\int_0^1\left|\frac{x^2y^2}{2}+\frac{x^3}{24}\right|_0^y dy$$

$$=\int_0^1\left(\frac{y^4}{2}+\frac{y^3}{24}\right)\mathrm{d}y$$

$$=\left|\frac{y^5}{10}+\frac{y^4}{96}\right|_0^1$$

$$=\frac{1}{10}+\frac{1}{96}$$

$$=\frac{53}{480}$$

(vi) $$P(X+Y\le 1)=\int_0^1\int_0^{1-y} f(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

$$=\int_0^1\int_0^{1-y}\left(xy^2+\frac{x^2}{8}\right)\mathrm{d}x\,\mathrm{d}y$$

$$=\int_0^1\left|\frac{x^2y^2}{2}+\frac{x^3}{24}\right|_0^{1-y}\mathrm{d}y$$

$$=\int_0^1\left\{\frac{(1-y)^2y^2}{2}+\frac{(1-y)^3}{24}\right\}\mathrm{d}y$$

$$=\int_0^1\left\{\frac{(1-2y+y^2)y^2}{2}+\frac{(1-y)^3}{24}\right\}\mathrm{d}y$$

$$=\int_0^1\left\{\frac{1}{2}(y^2-2y^3+y^4)+\frac{1}{24}(1-y)^3\right\}\mathrm{d}y$$

$$=\left|\frac{1}{2}\left(\frac{y^3}{3}-\frac{y^4}{2}+\frac{y^5}{5}\right)+\frac{1}{24}\frac{(1-y)^4}{(-4)}\right|_0^1$$

$$=\frac{1}{2}\left(\frac{1}{3}-\frac{1}{2}+\frac{1}{5}\right)+\frac{1}{24}\cdot\frac{1}{4}$$

$$=\frac{13}{480}$$

Example 14

The joint pdf of a two dimensional random variable (X, Y) is given by

$$f(x,y)=\frac{1}{2\pi a^2}e^{-\frac{x^2+y^2}{2a^2}},-\infty<x,y<\infty$$

Find $P(X^2+Y^2\le 4)$.

Solution

$$P(X^2+Y^2\le 4)=\iint_{x^2+y^2\le 4} f(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

$$=\iint_{x^2+y^2\le 4}\frac{1}{2\pi a^2}e^{-\frac{x^2+y^2}{2a^2}}\,\mathrm{d}x\,\mathrm{d}y$$

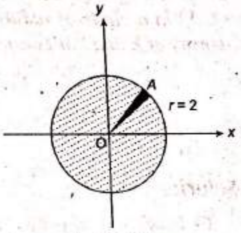


Fig. 2.8

The region of integration is the interior of the circle $x^2+y^2=4$.

Converting to polar coordinates by putting $x=\cos\theta, y=r\sin\theta, \mathrm{d}x\mathrm{d}y=r\mathrm{d}r\mathrm{d}\theta$, equation of the circle $x^2+y^2=4$ reduces to $r=2$.

In the region, along elementary radial strip *OA*,

Limits of $r: r=0$ to $r=2$

and in the region

Limits of θ: $\theta=0$ to $\theta=2\pi$

$$P(X^2+Y^2\le 4)=\int_0^{2\pi}\int_0^2\frac{1}{2\pi a^2}e^{-\frac{r^2}{2a^2}}\cdot r\mathrm{d}r\,\mathrm{d}\theta$$

$$=\frac{1}{2\pi}\int_0^{2\pi}\int_0^2 -e^{-\frac{r^2}{2a^2}}\left(-\frac{r}{a^2}\right)\mathrm{d}r\,\mathrm{d}\theta$$

$$=\frac{1}{2\pi}\int_0^{2\pi}\left|-e^{-\frac{r^2}{2a^2}}\right|_a^2\mathrm{d}\theta \qquad \left[\because \int_0 e^{f(x)}f'(x)dx=e^{f(x)}\right]$$

$$=\frac{1}{2\pi}\int_0^{2\pi}\left(-e^{-\frac{2}{a^2}}+1\right)\mathrm{d}\theta$$

$$=\frac{1}{2\pi}\left(1-e^{-\frac{2}{a^2}}\right)|\theta|_0^{2\pi}$$

$$=\frac{1}{2\pi}\left(1-e^{-\frac{2}{a^2}}\right)(2\pi)$$

$$=1-e^{-\frac{2}{a^2}}$$

Example 15

A gun is aimed at a certain point (origin of the co-ordinate system). Because of the random factors, the actual hit point can be any point (X, Y) in a circle of radius 'a' about the origin. Assume that the joint density of X and Y is constant in this circle and is given by

$$f(x,y) = c, \quad x^2 + y^2 \le a^2$$
$$= 0, \quad \text{otherwise}$$

Find (i) c (ii) $f_X(x)$.

Solution

(i) Since $f(x, y)$ is a probability density function,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1$$

$$\iint_{x^2+y^2\le a^2} c\,dx\,dy = 1$$

$$c\iint_{x^2+y^2\le a^2} dx\,dy = 1$$

$$c(\text{area of circle } x^2+y^2=a^2) = 1$$

$$c(\pi a^2) = 1$$

$$c = \frac{1}{\pi a^2}$$

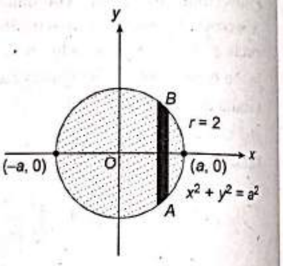


Fig. 2.9

(iii) The region of integration is the interior of the circle $x^2 + y^2 = a^2$.

In the region along the vertical strip AB, Limits of $y = -\sqrt{a^2 - x^2}$ to $y = \sqrt{a^2 - x^2}$ and x varies from $x = -a$ to $x = a$.

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$$

$$= \int_{-\sqrt{a^2-x^2}}^{\sqrt{a^2-x^2}} c\,dy$$

$$= c\left|y\right|_{-\sqrt{a^2-x^2}}^{\sqrt{a^2-x^2}}$$

$$= c\left(\sqrt{a^2-x^2} + \sqrt{a^2-x^2}\right)$$

$$= \frac{1}{\pi a^2}\left(2\sqrt{a^2-x^2}\right)$$

$$= \frac{2}{\pi a^2}\sqrt{a^2-x^2}, \qquad -a \le x \le a$$

EXERCISE 2.4

1. The joint pdf of a two dimensional random variable (X, Y) is given by

$$f(x,y) = 2, \quad 0 < x < 1, 0 < y < 1$$
$$= 0, \quad \text{otherwise}$$

Find (i) marginal density function (ii) conditional density function.

Ans.:
(i) $f(x) = 2x, \quad 0 < x < 1$
$= 0, \quad \text{otherwise}$
(ii) $f(y/x) = \frac{1}{x}, \quad 0 < x < 1$
$f(x/y) = \frac{1}{1-y}, 0 < y < 1$
$f(y) = 2(1-y), \quad 0 < y < 1$
$= 0, \quad \text{otherwise}$

2. The joint pdf of a two dimensional random variable (X, Y) is given by

$$f(x,y) = \frac{1}{ab}, \quad 0 < x < a, \quad 0 < y < b$$
$$= 0, \quad \text{otherwise}$$

Find the joint probability distribution function and marginal density functions. Are X and Y independent?

Ans.: $F(x,y) = \frac{xy}{ab}, \quad 0 < x \le a, \quad 0 < y \le b$
$= 0, \quad \text{otherwise}$
$f_X(x) = \frac{1}{a}, f_Y(y) = \frac{1}{b}$, Yes

3. The joint pdf of (X, Y) is given by $f(x,y) = k, 0 \le x < y \le 2$.
Find (i) k (ii) $f_X(x)$, $f_Y(y)$ (iii) $f(y/x)$, $f(x/y)$

Ans.: (i) $\frac{1}{2}$, (ii) $f_X(x) = \frac{1}{2}(2-x), 0 \le x \le 2; f_Y(y) = \frac{1}{2}y, 0 \le y \le 2$
(iii) $f(y/x) = \frac{1}{2-x}, x < y < 2; f(x/y) = \frac{1}{y}, 0 < x < y$

4. The joint pdf of (X, Y) is given by

$$f(x,y) = k(x^3y + xy^3), 0 \le x \le 2, 0 \le y \le 2$$

Find (i) k (ii) $f_X(x)$, $f_Y(y)$ (iii) $f(y/x)$, $f(x/y)$

$$\left[\begin{aligned}&\textbf{Ans.:} \text{ (i) } k=\frac{1}{16}, \text{ (ii) } f_X(x)=\frac{1}{8}(x^3+2x), 0\le x\le 2;\ f_Y(y)=\frac{1}{8}(y^3+2y), 0\le y\le 2\\ &\text{(iii) } f(y/x)=\frac{y(x^2+y^2)}{2(x^2+2)}, 0\le y\le 2;\ f(x/y)=\frac{x(x^2+y^2)}{2(y^2+2)}, 0\le x\le 2.\end{aligned}\right]$$

5. The joint pdf of (X, Y) is given by

$$\begin{aligned}f(x,y)&=\frac{1}{3}(3x^2+xy), && 0<x\le 1,\ 0<y\le 2\\ &=0, && \text{otherwise}\end{aligned}$$

Find $P(X+Y\ge 1)$.

$$\left[\textbf{Ans.:}\ \frac{65}{72}\right]$$

6. The joint pdf of (X, Y) is given by

$f(x,y) = k(6-x-y),\ 0<x<2,\ 2<y<4$

Find (i) k (ii) $P(X<1, Y<3)$, (iii) $P(X+Y<3)$, (iv) $P(X<1/y<3)$

$$\left[\textbf{Ans.:}\ \text{(i) } \frac{1}{8}, \text{(ii) } \frac{3}{8}, \text{(iii) } \frac{5}{24}, \text{(iv) } \frac{3}{5}\right]$$

7. The joint pdf of (X, Y) is given by

$$\begin{aligned}f(x,y)&=e^{-y}, && x>0,\ y>x\\ &=0, && \text{otherwise}\end{aligned}$$

Find (i) $P(X>1/Y<5)$ (ii) marginal distributions of X and Y.

$$\left[\textbf{Ans.:}\ \text{(i) } \frac{e^4-5}{e^5-6}, \text{(ii) } f_X(x)=e^{-x}, x>0;\ f_Y(y)=ye^{-y}, y>0\right]$$

8. The joint pdf of (X, Y) is given by

$$\begin{aligned}f(x,y)&=\frac{1}{12}e^{-\frac{x}{4}-\frac{y}{3}}, && x\ge 0, y\ge 0\\ &=0, && \text{otherwise}\end{aligned}$$

(i) Find conditional density functions of X and Y.

(ii) Are X, Y independent?

$$\left[\textbf{Ans.:}\ \text{(i) } f(y/x)=\frac{1}{3}e^{-\frac{y}{3}}, y\ge 0;\ f(x/y)=\frac{1}{4}e^{-\frac{x}{4}}, x\ge 0, \text{ (ii) Yes}\right]$$

9. The joint pdf of (X, Y) is given by

$$\begin{aligned}f(x,y)&=\frac{2}{(1+x+y)^3}, && x>0, y>0\\ &=0, && \text{otherwise}\end{aligned}$$

Find (i) $F(x, y)$ (ii) $f_X(x)$ (iii) $f(y/x)$.

$$\left[\begin{aligned}&\textbf{Ans.:} \text{ (i) } F(x,y)=1-\frac{1}{1+x}+\frac{1}{1+x+y}-\frac{1}{1+y}\\ &\text{(ii) } f_X(x)=\frac{1}{(1+x)^2}, \quad x>0\\ &\qquad\qquad =0, \quad \text{otherwise}\\ &\text{(iii) } f(y/x)=\frac{2(1+x)^2}{(1+x+y)^3}\end{aligned}\right]$$

10. The joint pdf of (X, Y) is given by

$$\begin{aligned}f(x,y)&=x^2+\frac{xy}{3}, && 0<x<1, 0<y<2\\ &=0, && \text{otherwise}\end{aligned}$$

Find (i) $P\left(X>\frac{1}{2}\right)$ (ii) $P(Y<X)$ (iii) $P\left(Y<\frac{1}{2}/X<\frac{1}{2}\right)$

$$\left[\textbf{Ans.:}\ \text{(i) } \frac{5}{6}, \text{ (ii) } \frac{7}{24}, \text{ (iii) } \frac{5}{32}\right]$$

11. The joint pdf of (X, Y) is given by

$$\begin{aligned}f(x,y)&=\frac{1}{4}(1+xy), && |x|<1,\ |y|<1\\ &=0, && \text{otherwise}\end{aligned}$$

Show that X and Y are independent.

12. The joint pdf of (X, Y) is given by $f(x,y)=Ae^{-|x|-2|y|}$. Show that X and Y are independent.

CHAPTER 3

Basic Statistics

Chapter Outline

3.1 Introduction
3.2 Measures of Central Tendency
3.3 Measures of Dispersion
3.4 Moments
3.5 Skewness
3.6 Kurtosis
3.7 Measures of Statistics for Continuous Random Variables
3.8 Expected Values of Two Dimensional Random Variables
3.9 Bounds on Probabilities
3.10 Chebyshev's Inequality

3.1 INTRODUCTION

A discrete random variable is described by probability function or probability mass function. Similarly, a continuous random variable is described by its probability density function. Instead of a function, a more compact description can be made by a few parameters, known as statistical measures, that are representative of the distribution. In descriptive statistics, statistical measures are used to summarize a set of observations in order to communicate the information as simply as possible. The observations are described in

(i) a measure of location or central tendency, such as arithmetic mean
(ii) a measure of statistical dispersion like standard deviation
(iii) a measure of the shape of the distribution like skewness or kurtosis
(iv) if more than one variable is measured, a measure of statistical dependence such as correlation coefficient.

3.2 MEASURES OF CENTRAL TENDENCY

In statistics, a central tendency or measure of central tendency is a central or typical value of a probability distribution. It is also called a center or location of the distribution. Measures of central tendency are often called averages. An average is a single value which can be taken as a representative of the whole distribution. There are five types of measures of central tendency or averages which are commonly used.

(i) Arithmetic mean or mean or expectation
(ii) Median
(iii) Mode
(iv) Geometric mean
(v) Harmonic mean

1. Mean The mean or average value (μ) of the probability distribution of a discrete random variable X is called as expectation and is denoted by $E(X)$.

$$\mu = E(X) = \sum_{i=1}^{\infty} x_i \, p(x_i) = \sum x \, p(x)$$

where $p(x)$ is the probability mass function of the discrete random variable X. Expectation of any function $\phi(x)$ of a random variable X is given by

$$E[\phi(x)] = \sum_{i=1}^{\infty} \phi(x_i) \, p(x_i) = \sum \phi(x) \, p(x)$$

Some important results on expectation:

(i) $E(X+k) = E(X) + k$
(ii) $E(aX \pm b) = aE(X) \pm b$
(iii) $E(X+Y) = E(X) + E(Y)$ provided $E(X)$ and $E(Y)$ exists.
(iv) $E(XY) = E(X)\,E(Y)$ if X and Y are two independent random variables.

2. Median The median is the point which divides the entire distribution into two equal parts. If X is a random variable, the value of $X = x$ for which the cumulative distribution function $F(x) = \frac{1}{2}$ is called the median of X. For a discrete random variable X, if there exits no x such that $F(x) = \frac{1}{2}$ then the median M of probability distribution is given by

$$M = \frac{1}{2}(x_k + x_{k+1})$$

where $F(x_k) < \frac{1}{2}$ and $F(x_{k+1}) > \frac{1}{2}$ and x_k and x_{k+1} are two consecutive values of X.

3. Mode The mode is the value of discrete random variable X for which the probability is maximum.

4. Geometric Mean The geometric mean G of a random variable X is defined by $\log G = E(\log X)$. The geometric mean of the probability distribution of a discrete random variable X is given by

$$\log G = \sum_{i=1}^{\infty} (\log x_i) p(x_i) = \sum (\log x) \, p(x)$$

where $p(x)$ is the probability mass function of the discrete random variable X.

5. Harmonic Mean The harmonic mean of a random variable X is defined by $\frac{1}{H} = E\left(\frac{1}{X}\right)$. The harmonic mean of the probability distribution of a discrete random variable X is given by

$$\frac{1}{H} = \sum_{i=1}^{\infty} \frac{1}{x_i} p(x_i) = \sum \frac{1}{x} p(x)$$

where $p(x)$ is the probability mass function of the discrete random variable X.

3.3 MEASURES OF DISPERSION

A measure of central tendency is a representative value of the random variable. But it is important to know how the values are clustered around or scattered away from the measure of central tendency. The property of the random variable or its distribution by which its values are clustered around or scattering away from the central value is called dispersion. There are three types of measures of dispersion which are commonly used.

(i) Quartile Deviation
(ii) Mean Deviation
(iii) Standard Deviation

1. Quartile Deviation Quartile deviation or semi-inter quartile range of the probability distribution of a discrete random variable X is given by

$$Q = \frac{1}{2}(Q_3 - Q_1)$$

where Q_1 and Q_3 are the first and third quartiles of the distribution respectively.

2. Mean Deviation Mean deviation of the probability distribution of a discrete random variable X is given by

$$\begin{aligned} MD &= E\{|X - \mu|\} \\ &= \sum_{i=1}^{\infty} |x_i - \mu| \, p(x_i) \\ &= \sum |x - \mu| \, p(x) \end{aligned}$$

where $p(x)$ is the probability mass function of the discrete random variable X.

3. Standard Deviation *Standard deviation* is the positive square root of the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean. It is denoted by the Greek letter σ.

$$\begin{aligned} SD = \sigma &= \sqrt{\sum_{i=1}^{\infty} x_i^2\ p(x_i) - \mu^2} \\ &= \sqrt{E(X^2) - \mu^2} \\ &= \sqrt{E(X^2) - [E(X)]^2} \end{aligned}$$

Variance Variance characterizes the variability in the distributions since two distributions with same mean can still have different dispersion of data about their means. Variance of the probability distribution of a discrete random variable X is given by

$$\begin{aligned} \text{Var}(X) = \sigma^2 &= E(X-\mu)^2 \\ &= E(X^2 - 2X\mu + \mu^2) \\ &= E(X^2) - E(2X\mu) + E(\mu^2) \\ &= E(X^2) - 2\mu\, E(X) + \mu^2 \qquad [\because E(\text{constant}) = (\text{constant})] \\ &= E(X^2) - 2\mu\mu + \mu^2 \\ &= E(X^2) - \mu^2 \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$

Some important results on variance:

(i) $\text{Var}(k) = 0$
(ii) $\text{Var}(kX) = k^2\,\text{Var}(X)$
(iii) $\text{Var}(X + k) = \text{Var}(X)$
(iv) $\text{Var}(aX + b) = a^2\,\text{Var}(X)$

Example 1

A random variable X has the following distribution:

X	1	2	3	4	5	6
$P(X = x)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$

Find (i) mean, (ii) variance, and (iii) $P(1 < X < 6)$.

Solution

(i) $\text{Mean} = \mu = \sum xp(x)$

$$= 1\left(\frac{1}{36}\right) + 2\left(\frac{3}{36}\right) + 3\left(\frac{5}{36}\right) + 4\left(\frac{7}{36}\right) + 5\left(\frac{9}{36}\right) + 6\left(\frac{11}{36}\right)$$

$$= \frac{161}{36}$$

$$= 4.47$$

(ii) $\text{Variance} = \sigma^2 = \sum x^2 p(x) - \mu^2$

$$= 1\left(\frac{1}{36}\right) + 4\left(\frac{3}{36}\right) + 9\left(\frac{5}{36}\right) + 16\left(\frac{7}{36}\right) + 25\left(\frac{9}{36}\right) + 36\left(\frac{11}{36}\right) - (4.47)^2$$

$$= \frac{791}{36} - 19.98$$

$$= 1.99$$

(iii) $P(1 < X < 6) = P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5)$

$$= \frac{3}{36} + \frac{5}{36} + \frac{7}{36} + \frac{9}{36}$$

$$= \frac{24}{36}$$

$$= 0.67$$

Example 2

The probability distribution of a random variable X is given below. Find (i) E(X), (ii) Var(X), *(iii) E(2X – 3), and (iv)* Var (2X – 3)

X	–2	–1	0	1	2
$P(X = x)$	0.2	0.1	0.3	0.3	0.1

Solution

(i) $E(X) = \Sigma x\, p(x)$

$$= -2(0.2) - 1(0.1) + 0 + (0.3) + 2(0.1)$$

$$= 0$$

(ii) $\text{Var}(X) = \Sigma x^2\ p(x) - [E(X)]^2$

$$= 4(0.2) + 1(0.1) + 0 + 1(0.3) + 4(0.1) - 0$$

$$= 1.6$$

(iii) $E(2X - 3) = 2E(X) - 3$

$$= 2(0) - 3$$

$$= -3$$

(iv) $\text{Var}(2X - 3) = (2)^2\,\text{Var}(X)$

$$= 4(1.6)$$

$$= 6.4$$

Example 3

Mean and standard deviation of a random variable X are 5 and 4 respectively. Find E(X^2) and standard deviation of (5 – 3X).

Solution

$$
\begin{aligned}
E(X) &= \mu = 5\\
SD &= \sigma = 4\\
\therefore \quad Var(X) &= \sigma^2 = 16\\
Var(X) &= E(X^2) - [E(X)]^2\\
16 &= E(X^2) - (5)^2\\
\therefore \quad E(X^2) &= 41\\
Var\,(5 - 3X) &= Var\,(5) - (-3)^2\,Var\,(X)\\
&= 0 + 9(16)\\
&= 144\\
SD\,(5 - 3X) &= \sqrt{Var\,(5 - 3X)}\\
&= \sqrt{144}\\
&= 12
\end{aligned}
$$

Example 4

A machine produces an average of 500 items during the first week of the month and on average of 400 items during the last week of the month, the probability for these being 0.68 and 0.32 respectively. Determine the expected value of the production. **[Summer 2015]**

Solution

Let X be the random variable which denotes the items produced by the machine. The probability distribution is

X	500	400
$P(X = x)$	0.68	0.32

Expected value of the production $E(X) = \sum x\, p(x)$

$$
\begin{aligned}
&= 500(0.68) + 400(0.32)\\
&= 468
\end{aligned}
$$

Example 5

The monthly demand for Allwyn watches is known to have the following probability distribution:

Demand (x)	1	2	3	4	5	6	7	8
Probability $p(x)$	0.08	0.12	0.19	0.24	0.16	0.10	0.07	0.04

Find the expected demand for watches. Also, compute the variance.

Solution

$$
\begin{aligned}
E(X) &= \Sigma x\, p(x)\\
&= 1(0.08) + 2(0.12) + 3(0.19) + 4(0.24) + 5(0.16)\\
&\quad + 6(0.10) + 7(0.07) + 8(0.04)\\
&= 4.06\\
Var(X) &= E(X^2) - [E(X)]^2\\
&= \Sigma x^2\, p(x) - [E(X)]^2\\
&= 1(0.08) + 4(0.12) + 9(0.19) + 16(0.24) + 25(0.16)\\
&\quad + 36(0.10) + 49(0.07) + 64(0.04) - (4.06)^2\\
&= 19.7 - 16.48\\
&= 3.21
\end{aligned}
$$

Example 6

A discrete random variable has the probability mass function given below:

X	–2	–1	0	1	2	3
$P(X = x)$	0.2	k	0.1	$2k$	0.1	$2k$

Find k, mean, and variance.

Solution

Since $P(X = x)$ is a probability mass function,

$$
\begin{aligned}
\sum P(X = x) &= 1\\
0.2 + k + 0.1 + 2k + 0.1 + 2k &= 1\\
5k + 0.4 &= 1\\
5k &= 0.6\\
k &= \frac{0.6}{5} = \frac{3}{25}
\end{aligned}
$$

Hence, the probability distribution is

X	–2	–1	0	1	2	3
$P(X = x)$	$\frac{2}{10}$	$\frac{3}{25}$	$\frac{1}{10}$	$\frac{6}{25}$	$\frac{1}{10}$	$\frac{6}{25}$

$$\text{Mean} = E(X) = \sum x\,p(x)$$

$$= (-2)\left(\frac{2}{10}\right) + (-1)\left(\frac{3}{25}\right) + 0 + 1\left(\frac{6}{25}\right) + 2\left(\frac{1}{10}\right) + 3\left(\frac{6}{25}\right)$$

$$= \frac{6}{25}$$

$$\text{Variance} = \text{Var}(X) = E(X^2) - [E(X)]^2$$

$$= \sum x^2\,p(x) - [E(X)]^2$$

$$= 4\left(\frac{2}{10}\right) + 1\left(\frac{3}{25}\right) + 0 + 1\left(\frac{6}{25}\right) + 4\left(\frac{1}{10}\right) + 9\left(\frac{6}{25}\right) - \left(\frac{6}{25}\right)^2$$

$$= \frac{73}{250} - \frac{36}{625}$$

$$= \frac{293}{625}$$

Example 7

A random variable X has the following probability function:

x	0	1	2	3	4	5	6	7
$p(x)$	0	k	$2k$	$2k$	$3k$	k^2	$2k^2$	$7k^2 + k$

(i) *Determine k.* (ii) *Evaluate* $P(X < 6)$, $P(X \geq 6)$, $P(0 < X < 5)$ and $P(0 \leq X \leq 4)$. (iii) *Determine the distribution function of X.* (iv) *Find the mean.* (v) *Find the variance.*

Solution

(i) Since $p(x)$ is a probability mass function,

$$\sum p(x) = 1$$

$$0 + k + 2k + 2k + 3k + k^2 + 2k^2 + 7k^2 + k = 1$$

$$10k^2 + 9k - 1 = 1$$

$$(10k - 1)(k + 1) = 0$$

$$k = \frac{1}{10} \text{ or } k = -1$$

$$k = \frac{1}{10} = 0.1 \left[\because\ p(x) \geq 0,\ k \neq -1\right]$$

Hence, the probability function is

X	0	1	2	3	4	5	6	7
$P(X = x)$	0	0.1	0.2	0.2	0.3	0.01	0.02	0.17

(ii) $P(X < 6) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5)$

$$= 0 + 0.1 + 0.2 + 0.2 + 0.3 + 0.01$$

$$= 0.81$$

$$P(X \geq 6) = 1 - P(X < 6)$$

$$= 1 - 0.81$$

$$= 0.19$$

$$P(0 < X < 5) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$$

$$= 0.1 + 0.2 + 0.2 + 0.3$$

$$= 0.8$$

$$P(0 \leq X \leq 4) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$$

$$= 0 + 0.1 + 0.2 + 0.2 + 0.3$$

$$= 0.8$$

(iii) Distribution function of X

x	$p(x)$	$F(x)$
0	0	0
1	0.1	0.1
2	0.2	0.3
3	0.2	0.5
4	0.3	0.8
5	0.01	0.81
6	0.02	0.83
7	0.17	1

(iv) $\mu = \sum xp(x)$

$$= 0 + 1(0.1) + 2(0.2) + 3(0.2) + 4(0.3) + 5(0.01) + 6(0.02) + 7(0.17)$$

$$= 3.66$$

(v) $\text{Var}(X) = \sigma^2 = \sum x^2\,p(x) - \mu^2$

$$= 0 + 1(0.1) + 4(0.2) + 9(0.2) + 16(0.3) + 25(0.01) + 36(0.02) + 49(0.17) - (3.66)^2$$

$$= 3.4044$$

Example 8

A fair dice is tossed. Let the random variable X denote the twice the number appearing on the dice. Write the probability distribution of X. Calculate mean and variance.

Solution

Let X be the random variable which denotes twice the number appearing on the dice.

(i) Probability distribution of X

x	2	4	6	8	10	12
$p(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

(ii) Mean $= \mu = \sum xp(x)$

$$= 2\left(\frac{1}{6}\right)+4\left(\frac{1}{6}\right)+6\left(\frac{1}{6}\right)+8\left(\frac{1}{6}\right)+10\left(\frac{1}{6}\right)+12\left(\frac{1}{6}\right)$$

$$= 7$$

(iii) Variance $= \sigma^2 = \sum x^2 p(x) - \mu^2$

$$= 4\left(\frac{1}{6}\right)+16\left(\frac{1}{6}\right)+36\left(\frac{1}{6}\right)+64\left(\frac{1}{6}\right)+100\left(\frac{1}{6}\right)+144\left(\frac{1}{6}\right)-(7)^2$$

$$= 11.67$$

Example 9

Two unbiased dice are thrown at random. Find the probability distribution of the sum of the numbers on them. Also, find mean and variance.

Solution

Let X be the random variable which denotes the sum of the numbers on two unbiased dice. The random variable X can take values 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12. The probability distribution is

X	2	3	4	5	6	7	8	9	10	11	12
$P(X = x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Mean $= \mu = \sum x\,p(x)$

$$= 2\left(\frac{1}{36}\right)+3\left(\frac{2}{36}\right)+4\left(\frac{3}{36}\right)+5\left(\frac{4}{36}\right)+6\left(\frac{5}{36}\right)+7\left(\frac{6}{36}\right)+8\left(\frac{5}{36}\right)$$
$$+9\left(\frac{4}{36}\right)+10\left(\frac{3}{36}\right)+11\left(\frac{2}{36}\right)+12\left(\frac{1}{36}\right)$$

$$= \frac{252}{36}$$

$$= 7$$

Variance $= \sigma^2 = \sum x^2\,p(x) - \mu^2$

$$= 4\left(\frac{1}{36}\right)+9\left(\frac{2}{36}\right)+16\left(\frac{3}{36}\right)+25\left(\frac{4}{36}\right)+36\left(\frac{5}{36}\right)$$
$$+49\left(\frac{6}{36}\right)+64\left(\frac{5}{36}\right)+81\left(\frac{4}{36}\right)+100\left(\frac{3}{36}\right)$$
$$+121\left(\frac{2}{36}\right)+144\left(\frac{1}{36}\right)-(7)^2$$

$$= \frac{1974}{36} - 49$$

$$= 5.83$$

Example 10

A sample of 3 items is selected at random from a box containing 10 items of which 4 are defective. Find the expected number of defective items.

Solution

Let X be the random variable which denotes the defective items.

Total number of items = 10

Number of good items = 6

Number of defective items = 4

$$P(X=0) = P(\text{no defective item}) = \frac{{}^{6}C_3}{{}^{10}C_3} = \frac{1}{6}$$

$$P(X=1) = P(\text{one defective item}) = \frac{{}^{6}C_2\,{}^{4}C_1}{{}^{10}C_3} = \frac{1}{2}$$

$$P(X=2) = P(\text{two defective items}) = \frac{{}^{6}C_1\,{}^{4}C_2}{{}^{10}C_3} = \frac{3}{10}$$

$$P(X=3) = P(\text{three defective items}) = \frac{{}^{4}C_3}{{}^{10}C_3} = \frac{1}{30}$$

Hence, the probability distribution is

X	0	1	2	3
$P(X=x)$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{3}{10}$	$\frac{1}{30}$

$$\text{Expected number of defective items} = E(X) = \sum x\,p(x)$$
$$= 0 + 1\left(\frac{1}{2}\right) + 2\left(\frac{3}{10}\right) + 3\left(\frac{1}{30}\right)$$
$$= 1.2$$

Example 11

A player tosses two fair coins. He wins ₹ 100 if a head appears and ₹ 200 if two heads appear. On the other hand, he loses ₹ 500 if no head appears. Determine the expected value of the game. Is the game favourable to the players?

Solution

Let X be the random variable which denotes the number of heads appearing in tosses of two fair coins.

$$S = \{HH, HT, TH, TT\}$$

$$p(x_1) = P(X=0) = P(\text{no heads}) = \frac{1}{4}$$

$$p(x_2) = P(X=1) = P(\text{one head}) = \frac{2}{4} = \frac{1}{2}$$

$$p(x_3) = P(X=2) = P(\text{two heads}) = \frac{1}{4}$$

Amount to be lost if no head appears $= x_1 = -$ ₹ 500

Amount to be won if one head appears $= x_2 =$ ₹ 100

Amount to be won if two heads appear $= x_3 =$ ₹ 200

$$\text{Expected value of the game} = \mu = \sum x\,p(x)$$
$$= x_1\,p(x_1) + x_2\,p(x_2) + x_3\,p(x_3)$$
$$= -500\left(\frac{1}{4}\right) + 100\left(\frac{1}{2}\right) + 200\left(\frac{1}{4}\right)$$
$$= ₹\,-25$$

Hence, the game is not favourable to the player.

Example 12

Amit plays a game of tossing a dice. If a number less than 3 appears, he gets ₹ a, otherwise he has to pay ₹ 10. If the game is fair, find a.

Solution

Let X be the random variable which denotes tossing of a dice.

Probability of getting a number less than 3, i.e., 1 or 2 $= p(x_1) = \frac{2}{6} = \frac{1}{3}$

Probability of getting number more than or equal to 3, i.e., 3, 4, 5, or 6 $= p(x_2) = \frac{4}{6} = \frac{2}{3}$

Amount to be received for number less than $3 = x_1 =$ ₹ a

Amount to be paid for numbers more than or equal to $3 = x_2 =$ ₹ -10

$$E(X) = \sum x\,p(x)$$
$$= x_1\,p(x_1) + x_2\,p(x_2)$$
$$= a\left(\frac{1}{3}\right) + (-10)\left(\frac{2}{3}\right)$$
$$= \frac{a}{3} - \frac{20}{3}$$

For a pair game, $E(x) = 0$.

$$\frac{a}{3} - \frac{20}{3} = 0$$
$$a = 20$$

Example 13

A man draws 2 balls from a bag containing 3 white and 5 black balls. If he is to receive ₹ 14 for every white ball which he draws and ₹ 7 for every black ball, what is his expectation?

Solution

Let X be the random variable which denotes the balls drawn from a bag. 2 balls drawn may be either (i) both white, or (ii) both black, or (iii) one white and one black.

Probability of drawing 2 white balls $= p(x_1) = \frac{{}^3C_2}{{}^8C_2} = \frac{3}{28}$

Probability of drawing 2 black balls $= p(x_2) = \frac{{}^5C_2}{{}^8C_2} = \frac{10}{28}$

Probability of drawing 1 white and 1 black ball = $p(x_3) = \dfrac{{}^3C_1\,{}^5C_1}{{}^8C_2} = \dfrac{15}{28}$

Amount to be received for 2 white balls = x_1 = ₹ 14 × 2 = ₹ 28

Amount to be received for 2 black balls = x_2 = ₹ 7 × 2 = ₹ 14

Amount to be received for 1 white and 1 black ball = x_3 = ₹ 14 + ₹ 7 = ₹ 21

$$\begin{aligned}\text{Expectation} = E(X) &= \sum x\,p(x)\\ &= x_1\,p(x_1) + x_2\,p(x_2) + x_3\,p(x_3)\\ &= 28\left(\frac{3}{28}\right) + 14\left(\frac{10}{28}\right) + 21\left(\frac{15}{28}\right)\\ &= ₹\,19.25\end{aligned}$$

Example 14

The probability that there is at least one error in an account statement prepared by A is 0.2 and for B and C, they are 0.25 and 0.4 respectively. A, B, and C prepared 10, 16, and 20 statements respectively. Find the expected number of correct statements in all.

Solution

Let $p(x_1)$, $p(x_2)$ and $p(x_3)$ be the probabilities of the events that there is no error in the account statements prepared by *A*, *B*, and *C* respectively.

$$\begin{aligned}p(x_1) &= 1 - (\text{Probability of at least one error in the account statement prepared by } A)\\ &= 1 - 0.2\\ &= 0.8\end{aligned}$$

Similarly, $p(x_2) = 1 - 0.25 = 0.75$

$p(x_3) = 1 - 0.4 = 0.6$

Also, $x_1 = 10, \quad x_2 = 16, \quad x_3 = 20$

$$\begin{aligned}\text{Expected number of correct statements} = E(X) &= \sum x\,p(x)\\ &= x_1\,p(x_1) + x_2\,p(x_2) + x_3\,p(x_3)\\ &= 10(0.8) + 16(0.75) + 20(0.6)\\ &= 32\end{aligned}$$

Example 15

A man has the choice of running either a hot-snack stall or an ice-cream stall at a seaside resort during the summer season. If it is a fairly cool summer, he should make ₹ 5000 by running the hot-snack stall, but if the summer is quite hot, he can only expect to make ₹ 1000. On the other hand, if he operates the ice-cream stall, his profit is estimated at ₹ 6500, if the summer is hot, but only ₹ 1000 if it is cool. There is a 40 percent chance of the summer being hot. Should he opt for running the hot-snack stall or the ice-cream stall?

Solution

Let *X* and *Y* be the random variables which denote the income from the hot-snack and ice-cream stalls respectively.

Probability of hot summer = $p_1 = 40\% = 0.4$

Probability of cool summer = $p_2 = 1 - p_1 = 1 - 0.4 = 0.6$

$x_1 = 1000, \quad x_2 = 5000, \quad y_1 = 6500, \quad y_2 = 1000$

$$\begin{aligned}\text{Expected income from hot-snack stall} &= E(X)\\ &= x_1\,p_1 + x_2\,p_2\\ &= 1000(0.4) + 5000(0.6)\\ &= ₹\,3400\end{aligned}$$

$$\begin{aligned}\text{Expected income from ice-cream stall} &= E(Y)\\ &= y_1\,p_1 + y_2\,p_2\\ &= 6500(0.4) + 1000(0.6)\\ &= ₹\,3200\end{aligned}$$

Hence, he should opt for running the hot-snack stall.

EXERCISE 3.1

1. The probability distribution of a random variable *X* is given by

X	–2	–1	0	1	2	3
P(X = x)	0.1	k	0.2	2k	0.3	k

Find *k*, the mean, and variance. [**Ans.:** 0.1, 0.8, 2.16]

2. Find the mean and variance of the following distribution:

X	4	5	6	8
P(X = x)	0.1	0.3	0.4	0.2

[**Ans.:** 5.9, 1.49]

3. Find the value of k from the following data:

X	0	10	15
$P(X = x)$	$\frac{k-6}{5}$	$\frac{2}{k}$	$\frac{14}{5k}$

Also, find the distribution function and expectation of X.

Ans.: 8,

X	0	10	15
$F(X)$	$\frac{2}{5}$	$\frac{13}{20}$	1

, $\frac{31}{4}$

4. For the following distribution,

X	−3	−2	−1	0	1	2
$P(X = x)$	0.01	0.1	0.2	0.3	0.2	0.15

Find (i) $P(X \geq 1)$, (ii) $P(X < 0)$, (iii) $E(X)$, and (iv) $\text{Var}(X)$

[**Ans.:** (i) 0.35 (ii) 0.35 (iii) 0.05 (iv) 1.8475]

5. A random variable X has the following probability function:

X	0	1	2	3	4	5	6	7	8
$P(X = x)$	$\frac{k}{45}$	$\frac{k}{15}$	$\frac{k}{9}$	$\frac{k}{5}$	$\frac{2k}{45}$	$\frac{6k}{45}$	$\frac{7k}{45}$	$\frac{8k}{45}$	$\frac{4k}{45}$

Determine (i) k, (ii) mean, (iii) variance, and (iv) SD.

[**Ans.:** (i) 1 (ii) 0.4622 (iii) 4.9971 (iv) 2.24]

6. A fair coin is tossed until a head or five tails appear. Find (i) discrete probability distribution, and (ii) mean of the distribution.

Ans.: (i)

X	1	2	3	4	5
$P(X = x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$

(ii) 1.9

7. Let X denotes the minimum of two numbers that appear when a pair of fair dice is thrown once. Determine (i) probability distribution, (ii) expectation, and (iii) variance.

Ans.: (i)

X	1	2	3	4	5	6
$P(X = x)$	$\frac{11}{36}$	$\frac{9}{36}$	$\frac{7}{36}$	$\frac{5}{36}$	$\frac{3}{36}$	$\frac{1}{36}$

(ii) 2.5278 (iii) 1.9713

8. For the following probability distribution,

X	−3	−2	−1	0	1	2	3
$P(X = x)$	0.001	0.01	0.1	?	0.1	0.01	0.001

Find (i) missing probability, (ii) mean, and (iii) variance.

[**Ans.:** (i) 0.778 (ii) 0.2 (iii) 0.258]

9. A discrete random variable can take all integer values from 1 to k each with the probability of $\frac{1}{k}$. Show that its mean and variance are $\frac{k+1}{2}$ and $\frac{k^2+1}{2}$ respectively.

10. An urn contains 6 white and 4 black balls; 3 balls are drawn without replacement. What is the expected number of black balls that will be obtained?

[**Ans.:** $\frac{6}{5}$]

11. A six-faced dice is tossed. If a prime number occurs, Anil wins that number of rupees but if a nonprime number occurs, he loses that number of rupees. Determine whether the game is favourable to the player.

[**Ans.:** The game is favourable to Anil]

12. A man runs an ice-cream parlour at a holiday resort. If the summer is mild, he can sell 2500 cups of ice cream; if it is hot, he can sell 4000 cups; if it is very hot, he can sell 5000 cups. It is known that for any year, the probability of summer to be mild is $\frac{1}{7}$ and to be hot is $\frac{4}{7}$. A cup of ice cream costs ₹ 2 and is sold for ₹ 3.50. What is his expected profit?

[**Ans.:** ₹ 6107.14]

13. A player tosses two fair coins. He wins ₹ 1 or ₹ 2 as 1 tail or 1 head appears. On the other hand, he loses ₹ 5 if no head appears. Find the expected gain or loss of the player.

[**Ans.:** Loss of ₹ 0.25]

14. A bag contains 2 white balls and 3 black balls. Four persons A, B, C, D in the order named each draws one ball and does not replace it. The first to draw a white ball receives ₹ 20. Determine their expectations.

[Ans.: ₹ 8, ₹ 6, ₹ 4, ₹ 2]

3.4 MOMENTS

Moment is the arithmetic mean of the various powers of the deviations of items from their assumed mean or actual mean. If the deviations of the items are taken from the arithmetic mean of the distribution, it is known as *central moment*. If the mean of the first power of deviations are taken, the first moment about the mean is obtained and is denoted by μ_1. The mean of the second power of the deviations gives the second moment about the mean and is denoted by μ_2. Similarly, the mean of the cubes of deviations gives third moment about the mean and is denoted by μ_3. The mean of the fourth power of the deviations from the mean gives the fourth moment about the mean and is denoted by μ_4. Thus, the mean of the r^{th} power of deviations gives the r^{th} moment about mean or r^{th} central moment and is denoted by μ_r.

3.4.1 Central Moments or Moments about Actual Mean

The moments about the mean value $\mu = E(X)$ are called central moments and denoted by μ_r.

$$\begin{aligned}\mu_r &= E\{(x-\mu)^r\}\\ &= \sum_{i=1}^{\infty}(x_i-\mu)^r\, p(x_i)\\ &= \sum (x-\mu)^r p(x)\end{aligned}$$

If frequency distribution is given and $n = \Sigma f$, then $p(x_i) = \dfrac{\sum f_i}{N}$

$$\mu_r = \frac{\sum_{i=1}^{\infty} f_i (x_i-\mu)^r}{N}$$

3.4.2 Properties of Central Moments

(i) The first moment about the mean is always zero, i.e., $\mu_1 = 0$.

(ii) The second moment about the mean measures variance, i.e.,

$$\mu_2 = \sigma^2 \quad \text{or} \quad \text{SD} = \sigma = \pm\sqrt{\mu_2}$$

(iii) The third moment about the mean measures skewness.

If $\mu_3 > 0$, the distribution is positively skewed.
If $\mu_3 < 0$, the distribution is negatively skewed.
If $\mu_3 = 0$, the distribution is symmetrical.

$$\text{Skewness } \beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

(iv) The fourth moment about the mean measures kurtosis. It gives information on the peakedness or height of the peak of a frequency distribution, i.e., whether it is more peaked or more flat topped than a normal curve.

$$\text{Kurtosis } \beta_2 = \frac{\mu_4}{\mu_2^2}$$

(v) In a symmetric distribution, all odd moments are zero, i.e., $\mu_1 = \mu_3 = \mu_5 = \cdots = \mu_{2r+1} = 0$.

3.4.3 Raw Moments or Moments about Arbitrary Origin

When the actual mean of a distribution is a fraction, it is tedious to calculate central moments. In such cases, moments about an arbitrary origin 'a' is calculated and then these moments are converted into the moments about actual mean. The moments about the arbitrary origin are known as raw moments and are denoted by μ_r'. Thus, μ_1' denotes the first moment about an arbitrary origin, μ_2' denotes the second moment about an arbitrary origin and so on.

$$\begin{aligned}\mu_r' &= E\{(X-a)^r\}\\ &= \sum_{i=1}^{\infty}(x_i-a)^r\, p(x_i)\\ &= \sum (x-a)^r p(x)\end{aligned}$$

If frequency distribution is given and $n = \Sigma f$, then $p(x_i) = \dfrac{\sum f_i}{N}$

$$\mu_r' = \frac{\sum_{i=1}^{\infty} f_i (x_i-a)^r}{N}$$

When $a = 0$, μ_r' is called r^{th} order simple moments.

$$\begin{aligned}\mu_r' &= E\{X^r\}\\ &= \sum_{i=1}^{\infty} x_i^r\, p(x_i)\end{aligned}$$

$$= \sum x^r p(x)$$

$$= \frac{\sum fx^r}{n}$$

3.4.4 Relation between Central Moments and Raw Moments

The moments about the actual mean, i.e., central moments and moments about the arbitrary origin, i.e., raw moments are related with each other by the following equations:

First central moment $\mu_1 = \mu_1' - \mu_1' = 0$

Second central moment $\mu_2 = \mu_2' - (\mu_1')^2$

Third central moment $\mu_3 = \mu_3' - 3\mu_2' \mu_1' + 2(\mu_1')^3$

Fourth central moment $\mu_4 = \mu_4' - 4\mu_3' \mu_1' + 6\mu_2' (\mu_1')^2 - 3(\mu_1')^4$

Similarly, the raw moments can be expressed in terms of central moments.

First raw moment $\mu_1' = \mu - a$

Second raw moment $\mu_2' = \mu_2 + (\mu_1')^2$

Third raw moment $\mu_3' = \mu_3 + 3\mu_2 \mu_1' + (\mu_1')^3$

Fourth raw moment $\mu_4' = \mu_4 + 4\mu_3 \mu_1' + 6\mu_2(\mu_1')^2 + (\mu_1')^4$

Example 1

Calculate the first four moments from the following data:

x	0	1	2	3	4	5	6	7	8
f	5	10	15	20	25	20	15	10	5

Also, calculate the values of β_1 and β_2.

Solution

$$N = \sum f = 125$$

$$\bar{x} = \frac{\sum fx}{N} = \frac{500}{125} = 4$$

x	f	fx	$x-\mu$	$f(x-\mu)$	$f(x-\mu)^2$	$f(x-\mu)^3$	$f(x-\mu)^4$
0	5	0	−4	20	80	−320	1280
1	10	10	−3	−30	90	−270	810
2	15	30	−2	−30	60	−120	240
3	20	60	−1	−20	20	−20	20
4	25	100	0	0	0	0	0
5	20	100	1	20	20	20	20
6	15	90	2	30	60	120	240
7	10	70	3	30	90	270	810
8	5	40	4	20	80	320	1280
	$\Sigma f = 125$	$\Sigma fx = 500$		$\Sigma f(x-\mu) = 0$	$\Sigma f(x-\mu)^2 = 500$	$\Sigma f(x-\mu)^3 = 0$	$\Sigma f(x-\mu)^4 = 4700$

Moments about the actual mean:

$$\mu_1 = \frac{\sum f(x-\mu)}{N} = \frac{0}{125} = 0$$

$$\mu_2 = \frac{\sum f(x-\mu)^2}{N} = \frac{500}{125} = 4$$

$$\mu_3 = \frac{\sum f(x-\mu)^3}{N} = \frac{0}{125} = 0$$

$$\mu_4 = \frac{\sum f(x-\mu)^4}{N} = \frac{4700}{125} = 37.6$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{0}{64} = 0$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{37.6}{16} = 2.35$$

Example 2

Calculate the first four moments of the following distribution about the mean:

x	0	1	2	3	4	5	6	7	8
f	1	8	28	56	70	56	28	8	1

Also, evaluate β_1 and β_2.

Solution

Let $a = 4$ be the arbitrary origin.

x	f	$x-a$	$f(x-a)$	$f(x-a)^2$	$f(x-a)^3$	$f(x-a)^4$
0	1	−4	−4	16	−64	256
1	8	−3	−24	72	−216	648
2	28	−2	−56	112	−224	448
3	56	−1	−56	56	−56	56
4	70	0	0	0	0	0
5	56	1	56	56	56	56
6	28	2	56	112	224	448
7	8	3	24	72	216	648
8	1	4	4	16	64	256
	$\sum f$ = 256		$\sum f(x-a)$ = 0	$\sum f(x-a)^2$ = 512	$\sum f(x-a)^3$ = 0	$\sum f(x-a)^4$ = 2816

$$N = \sum f = 256$$

Moments about the arbitrary origin:

$$\mu_1' = \frac{\sum f(x-a)}{N} = \frac{0}{256} = 0$$

$$\mu_2' = \frac{\sum f(x-a)^2}{N} = \frac{512}{256} = 2$$

$$\mu_3' = \frac{\sum f(x-a)^3}{N} = \frac{0}{256} = 0$$

$$\mu_4' = \frac{\sum f(x-a)^4}{N} = \frac{2816}{256} = 11$$

Moments about the actual mean:

$$\mu_1 = 0$$

$$\begin{aligned}\mu_2 &= \mu_2' - (\mu_1')^2 \\ &= 2 - 0 \\ &= 2\end{aligned}$$

$$\begin{aligned}\mu_3 &= \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3 \\ &= 0 - 3(2)(0) + 2(0)^3 \\ &= 0\end{aligned}$$

$$\begin{aligned}\mu_4 &= \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4 \\ &= 11 - 4(0)(0) + 6(2)(0)^2 - 3(0)^4 \\ &= 11\end{aligned}$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{11}{(2)^2} = 2.75$$

Example 3

The first four moments of distribution about $x = 2$ are 1, 2.5, 5.5, and 16. Calculate the four moments about μ.

Solution

$$\mu_1' = 1, \quad \mu_2' = 2.5, \; \mu_3' = 5.5, \quad \mu_4' = 16$$

Moments about the mean:

$$\mu_1 = 0$$

$$\begin{aligned}\mu_2 &= \mu_2' - (\mu_1')^2 \\ &= 2.5 - (1)^2 \\ &= 1.5\end{aligned}$$

$$\begin{aligned}\mu_3 &= \mu_3' - 3\mu_2\mu_1' + 2(\mu_1')^3 \\ &= 5.5 - 3(2.5)(1) + 2(1)^3 \\ &= 0\end{aligned}$$

$$\begin{aligned}\mu_4 &= \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4 \\ &= 16 - 4(5.5)(1) + 6(2.5)(1)^2 - 3(1)^4 \\ &= 6\end{aligned}$$

Example 4

The first three moments of a distribution about the value 2 of the variables are 1, 16, *and* −40. *Show that the mean* = 3, *variance* = 15 *and* $\mu_3 = -86$.

Solution

$$a = 2, \quad \mu_1' = 1, \quad \mu_2' = 16, \quad \mu_2' = 16, \quad \mu_3' = -40$$

$$\mu_1' = \mu - a$$

$$1 = \mu - 2$$

$$\therefore \quad \mu = 3$$

Mean = 3

$$\mu_2 = \mu_2' - (\mu_1')^2$$

$$= 16 - (1)^2$$

$$= 15$$

Variance = $\mu_2 = 15$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$$

$$= -40 - 3(16)(1) + 2(1)^3$$

$$= -86$$

EXERCISE 3.2

1. Calculate the first four moments about the mean from the following data:

x	1	2	3	4	5
f	2	3	5	4	1

[Ans.: 0, 1.262, 0.722, 3.795]

2. Calculate the first four moments about the mean and also the value of β_2 from the following table:

x	0	1	2	3	4	5	6	7	8
f	1	8	28	156	170	56	28	8	1

[Ans.: 0, 1.294, 0.642, 0.582, 3.93]

3. The first four moments of a distribution about the value 4 of the variables are 1, 4, 10, and 45. Show that the mean = 5, variance = 3, and $\mu_3 = 0$.
4. The first four central moments of a distribution are 0, 2.5, 0.7, and 18.75. Calculate β_1 and β_2.

[Ans.: 0.031, 3]

5. The values of μ_1, μ_2, μ_3 and μ_4 are 0, 9.2, 3.6, and 1.22 respectively. Find skewness and kurtosis of the distribution.

[Ans.: 0.129, 1.4]

6. The first four moments about the working mean 28.5 of a distribution are 0.294, 7.144, 14.409, and 454.98. Calculate the moments about the mean. Also, evaluate β_1 and β_2.

[Ans.: 28.794, 7.058, 36.151, 408.738, 3.717, 8.205]

3.5 SKEWNESS

Skewness is a measure that refers to the extent of symmetry or asymmetry in a distribution. A distribution is said to be *symmetrical* when its mean, median, and mode are equal, and the frequencies are symmetrically distributed about the mean. A symmetrical distribution when plotted on a graph will give a perfectly bell-shaped curve which is known as a *normal curve* (Fig. 3.1).

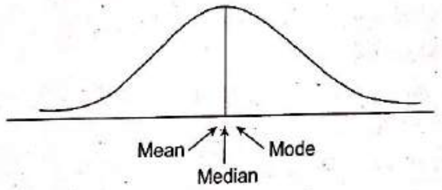
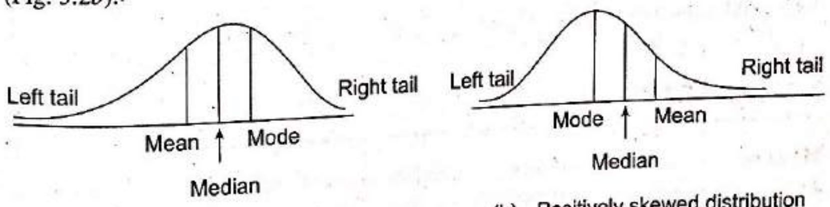


Fig. 3.1

A distribution is said to be asymmetrical or skewed when the mean, median, and mode are not equal, i.e., the mean median, and mode do not coincide. If the curve has a longer tail towards the left, it is said to be a negatively skewed distribution (Fig. 3.2*a*). If the curve has a longer tail towards the right, it is said to be positively skewed (Fig. 3.2*b*).



(a) Negatively skewed distribution

(b) Positively skewed distribution

Fig. 3.2

Skewness gives an idea of the nature and degree of concentration of observations about the mean.

3.5.1 Measures of Skewness

A measure of skewness gives the extent and direction of skewness of a distribution. These measures can be absolute or relative. The absolute measures are also known as measures of skewness.

Absolute skewness = Mean – Mode

If the value of the mean is greater than the mode, the skewness will be positive and if the value of the mean is less than the mode, the skewness will be negative.

The relative measures of skewness is called the *coefficient of skewness.*

3.5.2 Karl Pearson's Coefficient of Skewness

Karl Pearson's coefficient of skewness denoted by S_k, is given by

$$S_k = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

$$= \frac{\text{Mean} - \text{Mode}}{\sigma}$$

When the mode is ill-defined and the distribution is moderately skewed, the averages have the following relationship:

$$\text{Mode} = 3\,\text{Median} - 2\,\text{Mean}$$

$$S_k = \frac{\text{Mean} - (3\,\text{Median} - 2\,\text{Mean})}{\text{Standard Deviation}}$$

$$= \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

$$= \frac{3(\text{Mean} - \text{Median})}{\sigma}$$

The coefficient of skewness usually lies between –1 and 1.

For a positively skewed distribution, $S_k > 0$.

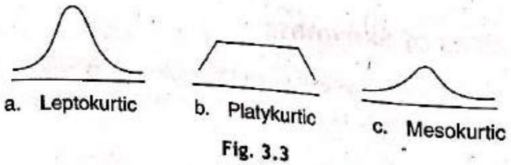
For a negatively skewed distribution, $S_k < 0$.

For a symmetrical distribution, $S_k = 0$.

3.6 KURTOSIS

Measures of central tendency, dispersion and skewness of a random variable cannot give a complete idea about the probability distribution. In order to analyse the probability distribution completely, another characteristic, Kurtosis is required. Kurtosis means the convexity of the probability curve of the distribution. It measures the degree of peakedness of distribution and is given by

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4}$$



a. Leptokurtic b. Platykurtic c. Mesokurtic

Fig. 3.3

The curves with $\beta_2 > 3$ is called Leptokurtic and those with $\beta_2 < 3$ are called platykurtic. The normal curve for which $\beta_2 = 3$ is called Mesokurtic.

As $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$ and $\beta_2 = \frac{\mu_4}{\mu_2^2}$ determine the shape of the probability curve, these are called Pearson's shape coefficients.

Example 1

From the marks scored by 100 *students in Section A and* 100 *students in Section B of a class, the following measures were obtained:*

Section *A*	$\mu_A = 55$	$\sigma_A = 15.4$	Mode = 58.72
Section *B*	$\mu_A = 53$	$\sigma_B = 15.4$	Mode = 48.83

Determine which distribution of marks is more skewed.

Solution

$$S_{k_A} = \frac{\text{Mean} - \text{Mode}}{\sigma_A} = \frac{55 - 58.72}{15.4} = -0.24$$

$$S_{k_B} = \frac{\text{Mean} - \text{Mode}}{\sigma_B} = \frac{53 - 48.83}{15.4} = 0.27$$

$$|0.27| > |-0.24|$$

Hence, the distribution of marks of Section *B* is more skewed.

Example 2

For a group of 10 *items,* $\sum x = 452, \sum x^2 = 24270$, *and mode* = 43.7. *Find Karl Pearson's coefficient of skewness.*

Solution

$n = 10, \quad \sum x = 452, \quad \sum x^2 = 24270, \quad \text{mode} = 43.7$

$$\mu = \frac{\Sigma x}{n} = \frac{452}{10} = 45.2$$

$$\sigma = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2}$$

$$= \sqrt{\frac{24270}{10} - \left(\frac{452}{10}\right)^2}$$

$$= 19.59$$

$$S_k = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

$$= \frac{45.2 - 43.7}{19.59}$$

$$= 0.077$$

Example 3

In a distribution, the mean = 65, median = 70, coefficient of skewness = –0.6. Find the mode and coefficient of variation.

Solution

$$\mu = 65, \quad \text{Median} = 70, \quad S_k = -0.6$$

$$\text{Mode} = 3\ \text{Median} - 2\ \text{Mean} = 3(70) - 2(65) = 80$$

$$S_k = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

$$-0.6 = \frac{65 - 80}{\sigma}$$

$$\therefore \quad \sigma = 25$$

$$\text{CV} = \frac{\sigma}{\bar{x}} \times 100 = \frac{25}{65} \times 100 = 38.64\%$$

Example 4

The following information was obtained from the records of a factory relating to wages:

Arithmetic mean = ₹ 56.8, Median = ₹ 59.5, *Standard deviation* = ₹ 12.4

Give the information about the distribution of wages.

Solution

$$\mu = 56.8, \quad \text{Median} = 59.5, \quad \sigma = 12.4$$

$$S_k = \frac{3(\text{Mean} - \text{Median})}{\sigma} = \frac{3(56.8 - 59.5)}{12.4} = -0.65$$

$$\text{Mode} = 3\ \text{Median} - 2\ \text{Mean} = 3(59.5) - 2(56.8) = 64.9$$

Hence, the maximum wages is ₹ 64.9.

There is a negative skewness in wages.

Example 5

For a moderately skewed distribution of retail price for men's shoes, it is found that the mean price is ₹ 20 and the median price is ₹ 17. If the coefficient of variation is 20%, find the Pearson's coefficient of skewness.

Solution

$$\mu = 20, \quad \text{Median} = 17, \quad \text{CV} = 20\%$$

$$\text{CV} = \frac{\sigma}{\bar{x}} \times 100$$

$$20 = \frac{\sigma}{20} \times 100$$

$$\therefore \quad \sigma = 4$$

$$S_k = \frac{3(\text{Mean} - \text{Median})}{\sigma} = \frac{3(20 - 17)}{4} = 2.25$$

Example 6

Find the mean, SD, quartiles, median and Karl Pearson's coefficient of skewness for the following probability distribution:

$X = x$	1	2	3	4	5	6	7	8
$p(x)$	0.008	0.032	0.142	0.216	0.240	0.206	0.143	0.013

Solution

(i) $$\text{Mean} = \mu = \sum x\, p(x) = 1(0.008) + 2(0.032) + 3(0.142) + 4(0.216) + 5(0.240) + 6(0.206) + 7(0.143) + 8(0.013)$$

$$= 4.903$$

(ii) $$\text{Var}(X) = \sigma^2 = \sum x^2 p(x) - \mu^2 = 1(0.008) + 4(0.032) + 9(0.142) + 16(0.216) + 25(0.240) + 36(0.206) + 49(0.143) + 64(0.013) - (4.903)^2$$

$$= 2.086$$

$$SD = \sqrt{\text{Var}(X)} = \sqrt{2.086} = 1.444$$

(iii) $$F(3) = 0.008 + 0.032 + 0.142 = 0.182 < 0.25$$

$$F(4) = 0.008 + 0.032 + 0.142 + 0.216 = 0.398 > 0.25$$

$$Q_1 = \frac{1}{3}(3 + 4) = 3.5$$

$$F(5) = 0.008 + 0.032 + 0.142 + 0.216 + 0.240 = 0.638 < 0.75$$

$$F(6) = 0.008 + 0.032 + 0.142 + 0.216 + 0.240 + 0.206 = 0.844 > 0.75$$

$$Q_2 = \frac{1}{2}(4+5) = 4.5$$

$$Q_3 = \frac{1}{2}(5+6) = 5.5$$

(iv) Median = $Q_2 = 4.5$

(v) Pearson's coefficient of skewness

$$S_k = \frac{\text{Mean} - \text{Median}}{\text{SD}} = \frac{4.903 - 4.5}{1.444} = 0.279$$

Example 7

Find the mean, median QD, MD, SD, β_1 and β_2 of the following probability distribution:

$X = x$	0	1	2	3	4	5	6	7	8
$p(x)$	0.004	0.036	0.1	0.232	0.280	0.204	0.112	0.028	0.004

Solution

(i) Mean = $\mu = \Sigma x\, p(x)$

$$= 0 + 1(0.036) + 2\,(0.1) + 3(0.232) + 4(0.280) + 5(0.204) + 6(0.112) + 7(0.028) + 8(0.04)$$

$$= 3.972$$

(ii) Median

$$F(3) = 0.004 + 0.036 + 0.1 + 0.232 = 0.372 < 0.5$$

$$F(4) = 0.004 + 0.036 + 0.1 + 0.232 + 0.280 = 0.652 > 0.5$$

$$\text{Median } M = \frac{1}{2}(3+4) = 3.5$$

(iii) Mode is the value of X for which $P(X = x)$ is maximum.

Mode = 4 [$\because P(X = 4) = 0.280$ is maximum probability]

(iv) Variance = $\sigma^2 = \sum x^2 p(x) - \mu^2$

$$= 0 + 1\,(0.036) + 4(0.1) + 9(0.232) + 16(0.28) + 25(0.204) + 36(0.112) + 49(0.028) + 64(0.004) - (3.972)^2$$

$$= 1.987$$

$$\text{SD} = \sqrt{\text{Var}(X)} = \sqrt{1.987} = 1.41$$

(v) $F(2) = 0.004 + 0.036 + 0.1 = 0.14 < 0.25$

$F(3) = 0.372 > 0.25$

$$Q_1 = \frac{1}{2}(2+3) = 2.5$$

$F(4) = 0.652 < 0.75$

$F(5) = 0.004 + 0.036 + 0.1 + 0.232 + 0.280 + 0.204 = 0.856 > 0.75$

$$Q_3 = \frac{1}{2}(4+5) = 4.5$$

$$QD = \frac{1}{2}(Q_3 - Q_1) = \frac{1}{2}(4.5 - 2.5) = 1$$

(vi) MD = $\sum |x - \mu|\, p(x)$

$$= 3.972(0.004) + 2.972(0.036) + 1.972(0.1) + 0.972(0.232) + 0.028\,(0.28) + 1.028(0.204) + 2.028(0.112) + 3.028(0.028) + 4.028(0.004)$$

$$= 1.091$$

(vii) $\mu_1' = \mu = 3.972$

$$\mu_2' = E(X^2) = \Sigma x^2 p(x)$$

$$= 0 + 1(0.036) + 4(0.1) + 9(0.232) + 16(0.280) + 25(0.204) + 36(0.112) + 49(0.028) + 64(0.004)$$

$$= 17.764$$

$$\mu_3' = E(X^3) = \Sigma x^3 p(x)$$

$$= 0 + 1(0.036) + 8\,(0.1) + 27(0.232) + 64(0.280) + 125(0.204) + 216(0.112) + 343(0.028) + 512(0.004)$$

$$= 86.364$$

$$\mu_4' = E(X^4) = \Sigma x^4 p(x)$$

$$= 0 + 1(0.036) + 16(0.1) + 81(0.232) + 256(0.280) + 625(0.204) + 1296(0.112) + 2401(0.028) + 4096(0.004)$$

$$= 448.372$$

$$\mu_2 = \sigma^2 = 1.987$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$$

$$= 86.364 - 3(17.764)\,(3.972) + 2(3.972)^3$$

$$= 0.019$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$$

$$= 448.372 - 4(86.364)\,(3.972) + 6(17.764)\,(3.972)^2 - 3(3.972)^4$$

$$= 11.053$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0.019)^2}{(1.987)^3} = 0.00005$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{11.053}{(1.987)^2} = 2.8$$

EXERCISE 3.3

1. Karl Pearson's measure of skewness of a distribution is 0.5. Its median and mode are respectively 42 and 36. Find the coefficient of variation.

   [Ans.: 40]

2. From the marks scored by 120 students in Section *A* and 120 students in Section *B* of a class, the following measures are obtained:

Section *A*	$\bar{x} = 46.83$	SD = 14.8	mode = 51.67
Section *B*	$\bar{x} = 47.83$	SD = 14.8	mode = 47.07

   Determine which distribution of marks is more skewed.

   [Ans.: Section *A*]

3. For a moderately skewed data, the arithmetic mean is 200, the coefficient of variation is 8, and Karl Pearson's coefficient of skewness is 0.3. Find the mode and median.

   [Ans.: 195.2, 198.4]

4. Karl Pearson's coefficient of skewness of a distribution is 0.32. Its standard deviation is 6.5 and the mean is 29.6. Find the mode and median for the distribution.

   [Ans.: 27.52, 28.9]

5. The median, mode and coefficient of skewness for a certain distribution are respectively 17.4, 15.3, and 0.35. Find the coefficient of variation.

   [Ans.: 48.78%]

6. In a distribution, mean = 65, median = 70, coefficient of skewness = –6. Find the mode and coefficient of variation.

   [Ans.: 80, 39.78%]

3.7 MEASURES OF STATISTICS FOR CONTINUOUS RANDOM VARIABLES

1. Mean The mean or average value (μ) of the probability distribution of a continuous random variable X is called the expectation and is denoted by $E(X)$.

$$\mu = E(X) = \int_{-\infty}^{\infty} x\, f(x)\,dx$$

Expectation of any function $\phi(x)$ of a continuous random variable X is given by

$$E[\phi(x)] = \int_{-\infty}^{\infty} \phi(x)\, f(x)\,dx$$

2. Median The median is the point which divides the entire distribution into two equal parts. In case of a continuous distribution, the median is the point which divides the total area into two equal parts. Thus, if a continuous random variable X is defined from a to b and M is the median,

$$\int_{a}^{M} f(x)\,dx = \int_{M}^{b} f(x)\,dx = \frac{1}{2}$$

By solving any one of this equation, the median is obtained.

3. Mode The mode is value of x for which $f(x)$ is maximum. Mode is given by

$$f'(x) = 0 \text{ and } f''(x) < 0 \text{ for } a < x < b$$

4. Geometric Mean The geometric mean of the probability distribution of a continuous random variable X is given by

$$\log G = \int_{-\infty}^{\infty} (\log x) f(x)\,dx$$

5. Harmonic Mean The harmonic mean of the probability distribution of a continuous random variable X is given by

$$\frac{1}{H} = \int_{-\infty}^{\infty} \frac{1}{x} f(x)\,dx$$

6. Quartile Deviation The r^{th} quartile of the probability distribution of a continuous random variable X and denoted by Q_r is given by

$$\int_{-\infty}^{Q_r} f(x)\,dx = \frac{r}{4},\ r = 1, 2, 3$$

Q_1 and Q_3 are called the first (lower) and the third (upper) quartiles respectively. Q_2 is the median (middle or second quartile). Quartile deviation or semi-interquartile range of the probability distribution of a continuous random variable X is given by

$$Q = \frac{1}{2}(Q_3 - Q_1)$$

7. Mean Deviation Mean deviation of the probability distribution of a continuous random variable X is given by

$$\text{MD} = \int_{-\infty}^{\infty} |x-\mu| f(x)\,dx$$

8. Standard Deviation The standard deviation of the probability distribution of a continuous random variable X is given by

$$\text{SD} = \sqrt{\text{Var}(X)} = \sigma$$

9. Variance The variance of the probability distribution of a continuous random variable X is given by

$$\text{Var}(X) = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx$$

$$= \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2$$

10. Moments Central moments or moments about actual mean of the probability distribution of a continuous random variable X is given by

$$\mu_r = \int_{-\infty}^{\infty} (x-\mu)^r f(x)\,dx$$

Raw moments or moments about arbitrary origin of the probability distribution of a continuous random variable X is given by

$$\mu_r' = \int_{-\infty}^{\infty} (x-a)^r f(x)\,dx$$

When $a = 0$, μ_r' is called r^{th} order simple moments.

$$\mu_r' = \int_{-\infty}^{\infty} x^r f(x)\,dx$$

11. Skewness Skewness of the probability distribution of a continuous random variable X is given by

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

12. Kurtosis Kurtosis of the probability distribution of a continuous random variable X is given by

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4}$$

Note The formulae of various measures of central tendency, dispersion, skewness and kurtosis of discrete probability distribution can be easily extended to the case of continuous probability distribution by simply replacing $p(x)$ by $f(x)dx$ and the summation by integration over the specified range of the variable X.

Example 1

For the continuous random variable having pdf

$$f(x) = 4x^3 \quad 0 \le x \le 1$$
$$= 0 \quad \textit{otherwise}$$

Find the mean and variance of X.

Solution

$$\text{Mean} = \mu = \int_{-\infty}^{\infty} x f(x)\,dx$$

$$= \int_{-\infty}^{0} x f(x)\,dx + \int_{0}^{1} x f(x)\,dx + \int_{1}^{\infty} x f(x)\,dx$$

$$= 0 + \int_{0}^{1} x(4x^3)\,dx + 0$$

$$= 4\int_{0}^{1} x^4\,dx$$

$$= 4\left|\frac{x^5}{5}\right|_0^1$$

$$= 4\left(\frac{1}{5} - 0\right)$$

$$= \frac{4}{5}$$

$$\text{Var}(X) = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2$$

$$= \int_{-\infty}^{0} x^2 f(x)\,dx + \int_{0}^{1} x^2 f(x)\,dx + \int_{1}^{\infty} x^2 f(x)\,dx - \mu^2$$

$$= 0 + \int_{0}^{1} x^2 (4x^3)\,dx + 0 - \left(\frac{4}{5}\right)^2$$

$$= 4\int_0^1 x^5\,dx - \frac{16}{25}$$

$$= 4\left|\frac{x^6}{6}\right|_0^1 - \frac{16}{25}$$

$$= \frac{4}{6} - \frac{16}{25}$$

$$= \frac{2}{75}$$

Example 2

For the triangular distribution

$$\begin{aligned} f(x) &= x & 0 < x \le 1 \\ &= 2 - x & 1 \le x \le 2 \\ &= 0 & \textit{otherwise} \end{aligned}$$

Find the mean and variance.

Solution

$$\mu = \int_{-\infty}^{\infty} x\,f(x)\,dx$$

$$= \int_{-\infty}^{0} x\,f(x)\,dx + \int_0^1 x\,f(x)\,dx + \int_1^2 x\,f(x) + \int_2^{\infty} x\,f(x)\,dx$$

$$= 0 + \int_0^1 x \cdot x\,dx + \int_1^2 x(2-x)\,dx + 0$$

$$= \int_0^1 x^2\,dx + \int_1^2 (2x - x^2)\,dx$$

$$= \left|\frac{x^3}{3}\right|_0^1 + \left|2\frac{x^2}{2} - \frac{x^3}{3}\right|_1^2$$

$$= \left(\frac{1}{3} - 0\right) + \left[\left(4 - \frac{8}{3}\right) - \left(1 - \frac{1}{3}\right)\right]$$

$$= \frac{1}{3} + \frac{4}{3} - \frac{2}{3}$$

$$= 1$$

$$\text{Var}(X) = \int_{-\infty}^{\infty} x^2\,f(x)\,dx - \mu^2$$

$$= \int_{-\infty}^{0} x^2 f(x)\,dx + \int_0^1 x^2\,f(x)\,dx + \int_1^2 x^2\,f(x)\,dx + \int_2^{\infty} x^2 f(x)\,dx - \mu^2$$

$$= 0 + \int_0^1 x^2 \cdot x\,dx + \int_1^2 x^2(2-x)\,dx + 0 - 1$$

$$= \int_0^1 x^3\,dx + \int_1^2 (2x^2 - x^3)\,dx - 1$$

$$= \left|\frac{x^4}{4}\right|_0^1 + \left|\frac{2x^3}{3} - \frac{x^4}{4}\right|_1^2 - 1$$

$$= \left(\frac{1}{4} - 0\right) + \left[\left(\frac{16}{3} - \frac{16}{4}\right) - \left(\frac{2}{3} - \frac{1}{4}\right)\right] - 1$$

$$= \frac{7}{6} - 1$$

$$= \frac{1}{6}$$

Example 3

If the probability density function of X is given by

$$f(x) = \begin{cases} \dfrac{x}{2} & 0 < x \le 1 \\ \dfrac{1}{2} & 1 < x \le 2 \\ \dfrac{3-x}{2} & 2 < x < 3 \\ 0 & \textit{otherwise} \end{cases}$$

Find the expected value of $f(x) = x^2 - 5x + 3$.

Solution

$$E[E\phi(x)] = \int_{-\infty}^{\infty} \phi(x) \cdot f(x)\,dx$$

$$E(x^2 - 5x + 3) = \int_{-\infty}^{\infty} (x^2 - 5x + 3)\,f(x)\,dx$$

$$= \int_0^1 (x^2 - 5x + 3)\frac{x}{2}\,dx + \int_1^2 (x^2 - 5x + 3)\frac{1}{2}\,dx + \int_2^3 (x^2 - 5x + 3)\left(\frac{3-x}{2}\right)dx$$

$$= \frac{1}{2}\int_0^1 (x^3 - 5x^2 + 3x)\,dx + \frac{1}{2}\int_1^2 (x^2 - 5x + 3)\,dx + \frac{1}{2}\int_2^3 (-x^3 + 8x^2 - 18x + 9)\,dx$$

$$= \frac{1}{2}\left|\frac{x^4}{4} - \frac{5x^3}{3} + \frac{3x^2}{2}\right|_0^1 + \frac{1}{2}\left|\frac{x^3}{3} - \frac{5x^2}{2} + 3x\right|_1^2 + \frac{1}{2}\left|-\frac{x^4}{4} + \frac{8x^3}{3} - \frac{18x^2}{2} + 9x\right|_2^3$$

$$= \frac{1}{2}\left(\frac{1}{4} - \frac{5}{3} + \frac{3}{2}\right) + \frac{1}{2}\left(\frac{8}{3} - 10 + 6 - \frac{1}{3} + \frac{5}{2} - 3\right) + \frac{1}{2}\left(-\frac{81}{4} + \frac{216}{3} - \frac{162}{2} + 27 + \frac{16}{4} - \frac{64}{3} + \frac{72}{2} - 18\right)$$

$$= \frac{1}{24} - \frac{13}{12} - \frac{19}{24}$$

$$= -\frac{11}{6}$$

Example 4

A continuous random variable has the probability density function

$$f(x) = kxe^{-\lambda x} \qquad x \ge 0, \lambda > 0$$
$$= 0 \qquad \textit{otherwise}$$

Determine (i) k, (ii) mean, and (iii) variance.

Solution

Since $f(x)$ is a probability density function,

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$\int_{-\infty}^{0} f(x)\,dx + \int_{0}^{\infty} f(x)\,dx = 1$$

$$0 + \int_0^{\infty} k\,x\,e^{-\lambda x}\,dx = 1$$

$$k\int_0^{\infty} x\,e^{-\lambda x}\,dx = 1$$

$$k\left|x\frac{e^{-\lambda x}}{-\lambda} - 1\frac{e^{-\lambda x}}{\lambda^2}\right|_0^{\infty} = 1$$

$$k\left[(0-0) - \left(0 - \frac{1}{\lambda^2}\right)\right] = 1$$

$$k = \lambda^2$$

Hence, $f(x) = \lambda^2 x e^{-\lambda x} \qquad x \ge 0, \lambda = 0$

$= 0 \qquad$ otherwise

(ii) Mean $= \mu = \int_{-\infty}^{\infty} x\,f(x)\,dx$

$$= \int_{-\infty}^{0} x\,f(x)\,dx + \int_0^{\infty} x\,f(x)\,dx$$

$$= 0 + \int_0^{\infty} x\,\lambda^2\,x e^{-\lambda x}\,dx$$

$$= \lambda^2 \int_0^{\infty} x^2\,e^{-\lambda x}\,dx$$

$$= \lambda^2\left|x^2\left(\frac{e^{-\lambda x}}{-\lambda}\right) - 2x\left(\frac{e^{-\lambda x}}{\lambda^2}\right) + 2\left(\frac{e^{-\lambda x}}{-\lambda^3}\right)\right|_0^{\infty}$$

$$= \lambda^2\left[(0-0+0) - \left(0 - 0 - \frac{2}{\lambda^3}\right)\right]$$

$$= \frac{2}{\lambda}$$

(iii) Variance $= \sigma^2 = \int_{-\infty}^{\infty} x^2\,f(x)\,dx - \mu^2$

$$= \int_{-\infty}^{0} x^2\,f(x)\,dx + \int_0^{\infty} x^2\,f(x)\,dx - \mu^2$$

$$= 0 + \int_0^{\infty} x^2\,\lambda^2\,x e^{-\lambda x}\,dx - \left(\frac{2}{\lambda}\right)^2$$

$$= \lambda^2 \int_0^{\infty} x^3\,e^{-\lambda x}\,dx - \frac{4}{\lambda^2}$$

$$= \lambda^2 \left| x^3 \left(\frac{e^{-\lambda x}}{-\lambda x} \right) - 3x^2 \left(\frac{e^{-\lambda x}}{\lambda^2} \right) + 6x \left(\frac{e^{-\lambda x}}{-\lambda^3} \right) - 6 \left(\frac{e^{-\lambda x}}{\lambda^4} \right) \right|_0^{\infty} - \frac{4}{\lambda^2}$$

$$= \lambda^2 \left[(0-0+0-0) - \left(0-0+0-\frac{6}{\lambda^4} \right) \right] - \frac{4}{\lambda^2}$$

$$= \frac{6}{\lambda^2} - \frac{4}{\lambda^2}$$

$$= \frac{2}{\lambda^2}$$

Example 5

The probability density f(x) of a continuous random variable is given by $f(x) = k\,e^{-|x|}$, $-\infty < x < \infty$ *(i) show that* $k = \frac{1}{2}$, *and (ii) find the mean and variance of the distribution. (iii) Also, find the probability that the variate lies between* 0 *and* 4.

Solution

(i) Since $f(x)$ is a probability density function,

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$\int_{-\infty}^{\infty} k e^{-|x|}\,dx = 1$$

$$k \int_{-\infty}^{\infty} e^{-|x|}\,dx = 1$$

$$2k \int_{0}^{\infty} e^{-|x|}\,dx = 1 \qquad \left[\because\ e^{-|x|} \text{ is an even function} \right]$$

$$2k \int_{0}^{\infty} e^{-x}\,dx = 1 \qquad \left[\because\ |x| = x \quad 0 \le x \le \infty \right]$$

$$2k \left| -e^{-x} \right|_0^{\infty} = 1$$

$$-2k\,(0-1) = 1$$

$$k = \frac{1}{2}$$

Hence, $f(x) = \frac{1}{2} e^{-|x|} \qquad -\infty < x < \infty$

(ii) $\mu = \int_{-\infty}^{\infty} x\, f(x)\,dx$

$$= \frac{1}{2} \int_{-\infty}^{\infty} x e^{-|x|}\,dx$$

$$= 0 \qquad \left[\because \text{ the integrand is an odd function} \right]$$

(iii) $\text{Var}(X) = \sigma^2 = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2$

$$= \frac{1}{2} \int_{-\infty}^{\infty} x^2 e^{-|x|}\,dx - 0$$

$$= 2\left(\frac{1}{2} \right) \int_{0}^{\infty} x^2 e^{-|x|}\,dx \qquad \left[\because \text{ the integrand is an even function} \right]$$

$$= \int_{0}^{\infty} x^2 e^{-|x|}\,dx$$

$$= \left| x^2 \frac{e^{-x}}{-1} - 2x \frac{e^{-x}}{1} + 2 \frac{e^{-x}}{-1} \right|_0^{\infty}$$

$$= 0 - (-2)$$

$$= 2$$

(iii) Probability that the variate lies between 0 and 4

$$P(0 < X < 4) = \int_0^4 f(x)\,dx$$

$$= \frac{1}{2} \int_0^4 e^{-|x|}\,dx$$

$$= \frac{1}{2} \int_0^4 e^{-x}\,dx \qquad \left[\because\ |x| = x \quad 0 < x < 4 \right]$$

$$= -\frac{1}{2} \left| e^{-x} \right|_0^4$$

$$= -\frac{1}{2} (e^{-4} - 1)$$

$$= 0.4908$$

Example 6

The daily consumption of electric power is a random variable X with probability density function

$$
\begin{aligned}
f(x) &= k\,x\,e^{-\frac{x}{3}} & x &> 0\\
&= 0 & x &\le 0
\end{aligned}
$$

Find the value of k, the expectation of X, and the probability that on a given day, the electric consumption is more than the expected value.

Solution

Since $f(x)$ is a probability density function,

$$
\begin{aligned}
\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x &= 1\\
\int_{-\infty}^{0} f(x)\,\mathrm{d}x + \int_{0}^{\infty} f(x)\,\mathrm{d}x &= 1\\
0 + \int_{0}^{\infty} k\,x\,e^{-\frac{x}{3}}\,\mathrm{d}x &= 1\\
k\left| x\left(\frac{e^{-\frac{x}{3}}}{-\frac{1}{3}}\right) - (1)\left(\frac{e^{-\frac{x}{3}}}{\frac{1}{9}}\right)\right|_{0}^{\infty} &= 1\\
k\left[(0-0)-(0-9)\right] &= 1\\
9k &= 1\\
k &= \frac{1}{9}
\end{aligned}
$$

Hence,
$$
\begin{aligned}
f(x) &= \frac{1}{9}xe^{-\frac{x}{3}} & x &> 0\\
&= 0 & x &\le 0
\end{aligned}
$$

$$
\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} x\,f(x)\,\mathrm{d}x\\
&= \int_{-\infty}^{0} x\,f(x)\,\mathrm{d}x + \int_{0}^{\infty} x\,f(x)\,\mathrm{d}x\\
&= 0 + \int_{0}^{\infty} x\cdot\frac{1}{9}\,x\,e^{-\frac{x}{3}}\,\mathrm{d}x\\
&= \frac{1}{9}\int_{0}^{\infty} x^{2}\,e^{-\frac{x}{3}}\,\mathrm{d}x
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{9}\left| x^{2}\left(\frac{e^{-\frac{x}{3}}}{-\frac{1}{3}}\right) - 2x\left(\frac{e^{-\frac{x}{3}}}{\frac{1}{9}}\right) + 2\left(\frac{e^{-\frac{x}{3}}}{-\frac{1}{27}}\right)\right|_{0}^{\infty}\\
&= \frac{1}{9}(0-0+0+54)\\
&= 6
\end{aligned}
$$

$$
\begin{aligned}
P(X>6) &= \int_{0}^{6} f(x)\,\mathrm{d}x\\
&= \int_{0}^{6} \frac{1}{9}xe^{-\frac{x}{3}}\,\mathrm{d}x\\
&= \frac{1}{9}\int_{0}^{6} xe^{-\frac{x}{3}}\,\mathrm{d}x\\
&= \frac{1}{9}\left| x\left(\frac{e^{-\frac{x}{3}}}{-\frac{1}{3}}\right) - 1\left(\frac{e^{-\frac{x}{3}}}{\frac{1}{9}}\right)\right|_{0}^{\infty}\\
&= \frac{1}{9}\left[(0-0)-\left(-18e^{-2}-9e^{-2}\right)\right]\\
&= 3\,e^{-2}\\
&= 0.406
\end{aligned}
$$

Example 7

Let X be a random variable with E(X) = 10 and Var(X) = 25. Find the positive values of a and b such that Y = aX − b has an expectation of 0 and a variance of 1.

Solution

$$
\begin{aligned}
E(Y) &= E(aX-b)\\
0 &= aE(X)-b\\
&= a(10)-b\\
10a-b &= 0\\
\mathrm{Var}(Y) &= \mathrm{Var}(aX-b)\\
1 &= a^{2}\,\mathrm{Var}(X)\\
&= a^{2}(25)
\end{aligned}
$$

$$25a^2 = 1$$

$$a = \frac{1}{5}$$

$$b = 2$$

Example 8

A continuous random variable X is distributed over the interval [0, 1] *with pdf* $f(x) = ax^2 + bx$, *where a, b are constants. If the mean of X is* 0.5, *find the values of a and b.*

Solution

Since $f(x)$ is probability density function,

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$\int_{-\infty}^{0} f(x)\,dx + \int_{0}^{1} f(x)\,dx + \int_{1}^{\infty} f(x)\,dx = 1$$

$$0 + \int_{0}^{1} (ax^2 + bx)\,dx + 0 = 1$$

$$\left| \frac{ax^3}{3} + \frac{bx^2}{2} \right|_0^1 = 1$$

$$\frac{a}{3} + \frac{b}{2} = 1$$

$$2a + 3b = 6 \qquad \text{...(1)}$$

Also, $\mu = 0.5$

$$\int_{0}^{1} x\,f(x)\,dx = 0.5$$

$$\int_{0}^{1} x\,(ax^2 + bx)\,dx = 0.5$$

$$\int_{0}^{1} (ax^3 + bx^2)\,dx = 0.5$$

$$\left| \frac{ax^4}{4} + \frac{bx^3}{3} \right|_0^1 = 0.5$$

$$\frac{a}{4} + \frac{b}{3} = 0.5$$

$$3a + 4b = 6 \qquad \text{...(2)}$$

Solving Eqs (1) and (2),

$$a = -6, \qquad b = 6$$

Example 9

A continuous random variable X has the pdf defined by $f(x) = A + Bx$, $0 \le x \le 1$. *If the mean of the distribution is* $\frac{1}{3}$, *find A and B.*

Solution

Since $f(x)$ is a probability density function,

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$\int_{-\infty}^{0} f(x)\,dx + \int_{0}^{1} f(x)\,dx + \int_{1}^{\infty} f(x)\,dx = 1$$

$$0 + \int_{0}^{1} (A + Bx)\,dx + 0 = 1$$

$$\left| Ax + \frac{Bx^2}{2} \right|_0^1 = 1$$

$$A + \frac{B}{2} = 1 \qquad \text{...(1)}$$

Also,

$$\mu = \frac{1}{3}$$

$$\int^{\infty} x\,f(x)\,dx = \frac{1}{3}$$

$$\int_{-\infty}^{0} x\,f(x)\,dx + \int_{0}^{1} x\,f(x)\,dx = \frac{1}{3}$$

$$0 + \int_{0}^{1} x\,(A + Bx)\,dx = \frac{1}{3}$$

$$\int_{0}^{1} (Ax + Bx^2)\,dx = \frac{1}{3}$$

$$\left| \frac{Ax^2}{2} + \frac{Bx^3}{3} \right|_0^1 = \frac{1}{3}$$

$$\frac{A}{2} + \frac{B}{3} = \frac{1}{3}$$

$$3A + 2B = 2 \qquad \text{...(2)}$$

Solving Eqs (1) and (2),

$$A = 2, \quad B = -2$$

Example 10

A continuous random variable has probability density function $f(x) = 6(x - x^2) \quad 0 \le x \le 1.$

Find the (i) mean, (ii) variance, (iii) median, and (iv) mode.

Solution

(i) $\mu = \int_{-\infty}^{\infty} x\, f(x)\, dx$

$$= \int_{-\infty}^{0} x\, f(x)\, dx + \int_{0}^{1} x\, f(x)\, dx + \int_{1}^{\infty} x\, f(x)\, dx$$

$$= 0 + \int_{0}^{1} x\, 6(x - x^2)\, dx + 0$$

$$= 6\int_{0}^{1} (x^2 - x^3)\, dx$$

$$= 6\left|\frac{x^3}{3} - \frac{x^4}{4}\right|_0^1$$

$$= 6\left(\frac{1}{3} - \frac{1}{4}\right)$$

$$= \frac{1}{2}$$

(ii) $\text{Var}(X) = \int_{-\infty}^{\infty} x^2 f(x)\, dx - \mu^2$

$$= \int_{-\infty}^{0} x^2 f(x)\, dx + \int_{0}^{1} x^2\, f(x)\, dx + \int_{1}^{\infty} x^2 f(x)\, dx - \mu^2$$

$$= 0 + \int_0^1 x^2\, 6(x - x^2)\, dx + 0 - \frac{1}{4}$$

$$= 6\int_0^1 (x^3 - x^4)\, dx - \frac{1}{4}$$

$$= 6\left|\frac{x^4}{4} - \frac{x^5}{5}\right|_0^1 - \frac{1}{4}$$

$$= 6\left(\frac{1}{4} - \frac{1}{5}\right) - \frac{1}{4}$$

$$= \frac{6}{20} - \frac{1}{4}$$

$$= \frac{1}{20}$$

(iii) $\int_{a}^{M} f(x)\, dx = \int_{M}^{b} f(x)\, dx = \frac{1}{2}$

$$\int_{0}^{M} 6(x - x^2)\, dx = \frac{1}{2}$$

$$6\left|\frac{x^2}{2} - \frac{x^3}{3}\right|_0^M = \frac{1}{2}$$

$$6\left(\frac{M^2}{2} - \frac{M^3}{3}\right) = \frac{1}{2}$$

$$3M^2 - 2M^3 = \frac{1}{2}$$

$$4M^3 - 6M^2 + 1 = 0$$

$$(2M - 1)(2M^2 - 2M - 1) = 0$$

$$M = \frac{1}{2} \quad \text{or} \quad M = \frac{1 \pm \sqrt{3}}{2}$$

$$M = \frac{1}{2} \text{ lies in } (0, 1)$$

Hence, median $M = \frac{1}{2}$

(iv) Mode is the value of x for which $f(x)$ is maximum. For $f(x)$ to be maximum, $f'(x) = 0$ and $f''(x) < 0$.

$$f'(x) = 0$$

$$6(1 - 2x) = 0$$

$$x = \frac{1}{2}$$

$$f''(x) = -12x$$

At $x = \frac{1}{2}$, $f''(x) = -12 < 0$

Hence, $f(x)$ is maximum at $x = \frac{1}{2}$.

Mode $= \frac{1}{2}$

Example 11

The probability density function of a random variable X is

$$f(x)=\frac{1}{2}\sin x \quad 0\le x\le \pi$$
$$=0 \qquad \textit{otherwise}$$

Find the mean, mode, and median of the distribution and also, find the probability between 0 and $\frac{\pi}{2}$.

Solution

(i) $\mu=\int_{-\infty}^{\infty} f(x)\,dx$

$$=\int_{-\infty}^{0} x\,f(x)\,dx+\int_{0}^{\pi} x\,f(x)\,dx+\int_{\pi}^{\infty} x\,f(x)\,dx$$
$$=0+\int_{0}^{\pi} x\left(\frac{1}{2}\sin x\right)dx+0$$
$$=\frac{1}{2}\int_{0}^{\pi} x\sin x\,dx$$
$$=\frac{1}{2}\left|-x\cos x+\sin x\right|_0^{\pi}$$
$$=\frac{\pi}{2}$$

(ii) Mode is the value of x for which $f(x)$ is maximum. For $f(x)$ to be maximum, $f'(x)=0$ and $f''(x)<0$.

$$f'(x)=0$$
$$\cos x=0$$
$$x=\frac{\pi}{2}$$
$$f''(x)=-\frac{1}{2}\sin x$$

At $x=\frac{\pi}{2}, f''(x)=-\frac{1}{2}<0$

Hence, $f(x)$ is maximum of $x=\frac{\pi}{2}$.

Mode $=\frac{\pi}{2}$

(iii) $\int_{a}^{M} f(x)\,dx=\int_{M}^{b} f(x)\,dx=\frac{1}{2}$

$$\int_{0}^{M}\frac{1}{2}\sin x\,dx=\int_{M}^{\pi}\frac{1}{2}\sin x\,dx=\frac{1}{2}$$
$$\int_{0}^{M}\frac{1}{2}\sin x\,dx=\frac{1}{2}$$
$$-\frac{1}{2}\left|\cos x\right|_0^{M}=\frac{1}{2}$$
$$-\frac{1}{2}(\cos M-1)=\frac{1}{2}$$
$$1-\cos M=0$$
$$\cos M=0$$
$$M=\frac{\pi}{2}$$

Hence, median $M=\frac{\pi}{2}$

(iv) $P\left(0<X<\frac{\pi}{2}\right)=\int_0^{\frac{\pi}{2}} f(x)\,dx$

$$=\int_0^{\frac{\pi}{2}}\frac{1}{2}\sin x\,dx$$
$$=-\frac{1}{2}\left|\cos x\right|_0^{\frac{\pi}{2}}$$
$$=-\frac{1}{2}(0-1)$$
$$=\frac{1}{2}$$

Example 12

The cumulative distribution function of a continuous random variable X is $F(x)=1-e^{-2x} \quad x\ge 0$

$$=0 \qquad x<0$$

Find the (i) the probability density function, (ii) mean, and (iii) variance.

Solution

(i) $f(x)=\dfrac{\mathrm{d}}{\mathrm{d}x}F(x)$

$$\begin{aligned} f(x) &= \frac{1}{2}e^{-2x} && x\ge 0\\ &= 0 && x<0 \end{aligned}$$

(ii) $\mu=\int_{-\infty}^{\infty} x\, f(x)\,\mathrm{d}x$

$$\begin{aligned} &= \int_{-\infty}^{0} x\, f(x)\,\mathrm{d}x+\int_{0}^{\infty} x\, f(x)\,\mathrm{d}x\\ &= 0+\int_{0}^{\infty} x\cdot\frac{1}{2}e^{-2x}\,\mathrm{d}x\\ &= \frac{1}{2}\int_{0}^{\infty} x\, e^{-2x}\,\mathrm{d}x\\ &= \frac{1}{2}\left|x\left(\frac{e^{-2x}}{-2}\right)-1\left(\frac{e^{-2x}}{4}\right)\right|_{0}^{\infty}\\ &= \frac{1}{2}\left[(0-0)-\left(0-\frac{1}{4}\right)\right]\\ &= \frac{1}{8} \end{aligned}$$

(iii) $\mathrm{Var}(X)=\int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x-\mu^2$

$$\begin{aligned} &= \int_{-\infty}^{0} x^2 f(x)\,\mathrm{d}x+\int_{0}^{\infty} x^2 f(x)\,\mathrm{d}x-\mu^2\\ &= 0+\int_{0}^{\infty} x^2\cdot\frac{1}{2}e^{-2x}\,\mathrm{d}x-\left(\frac{1}{8}\right)^2\\ &= \frac{1}{2}\int_{0}^{\infty} x^2 e^{-2x}\mathrm{d}x-\frac{1}{64}\\ &= \frac{1}{2}\left|x^2\left(\frac{e^{-2x}}{-2}\right)-2x\left(\frac{e^{-2x}}{4}\right)+2\left(\frac{e^{-2x}}{-8}\right)\right|_{0}^{\infty}-\frac{1}{64}\\ &= \frac{1}{2}\left[(0-0-0)-\left(0-0-\frac{1}{4}\right)\right]-\frac{1}{64}\\ &= \frac{1}{8}-\frac{1}{64}\\ &= \frac{7}{64} \end{aligned}$$

Example 13

A continuous random variable X has the distribution function

$$\begin{aligned} F(x) &= 0 && x\le 1\\ &= k(x-1)^4 && 1<x\le 3\\ &= 1 && x>3 \end{aligned}$$

Determine (i) $f(x)$, (ii) k, and (iii) mean.

Solution

(i) $f(x)=\dfrac{\mathrm{d}}{\mathrm{d}x}F(x)$

$$\begin{aligned} f(x) &= 0 && x\le 1\\ &= 4k(x-1)^3 && 1<x\le 3\\ &= 0 && x>3 \end{aligned}$$

(ii) Since $f(x)$ is a probability density function,

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)\mathrm{d}x &= 1\\ \int_{-\infty}^{1} f(x)\mathrm{d}x+\int_{1}^{3} f(x)\mathrm{d}x+\int_{3}^{\infty} f(x)\mathrm{d}x &= 1\\ 0+\int_{1}^{3} 4k(x-1)^3\,\mathrm{d}x+0 &= 1\\ 4k\left|\frac{(x-1)^4}{4}\right|_{1}^{3} &= 1\\ k(16-0) &= 1\\ k &= \frac{1}{16} \end{aligned}$$

Hence,
$$\begin{aligned} f(x) &= 0 && x\le 1\\ &= \frac{1}{4}(x-1)^3 && 1<x\le 3\\ &= 0 && x>3 \end{aligned}$$

(iii) $\mu=\int_{-\infty}^{\infty} x\, f(x)\,\mathrm{d}x$

$$\begin{aligned} &= \int_{-\infty}^{1} x\, f(x)\,\mathrm{d}x+\int_{1}^{3} x\, f(x)\,\mathrm{d}x+\int_{3}^{\infty} x\, f(x)\,\mathrm{d}x\\ &= 0+\int_{1}^{3} x\cdot\frac{1}{4}(x-1)^3\,\mathrm{d}x+0\\ &= \frac{1}{4}\int_{1}^{3} x(x-1)^3\,\mathrm{d}x \end{aligned}$$

$$= \frac{1}{4}\int_0^2 (t+1)\,t^3\,dt \qquad \begin{bmatrix}\text{Putting } x-1=t\\ \text{When } x=1, t=0\\ \text{When } x=3, t=2\end{bmatrix}$$

$$= \frac{1}{4}\int_0^2 (t^4+t^3)\,dt$$

$$= \frac{1}{4}\left|\frac{t^5}{5}+\frac{t^4}{4}\right|_0^2$$

$$= \frac{1}{4}\left[\left(\frac{2^5}{5}+\frac{2^4}{4}\right)-(0)\right]$$

$$= 2.6$$

Example 14

If the density function of a random variable X is given by

$$f(x) = kx\,(1-x), \quad 0 \le x \le 1,$$

find (i) AM, (ii) HM,'(iii) Median, (iv) Mode, (v) SD, (vi) MD about the mean.

Solution

(i) Since $f(x)$ is a probability density function,

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$$

$$\int_0^1 kx(1-x)\mathrm{d}x = 1$$

$$k\int_0^1 (x-x^2)\,\mathrm{d}x = 1$$

$$k\left|\frac{x^2}{2}-\frac{x^3}{3}\right|_0^1 = 1$$

$$k\left(\frac{1}{2}-\frac{1}{3}\right) = 1$$

$$k = 6$$

Hence, $f(x) = 6x(1-x), \quad 0 \le x \le 1$

(ii) $\text{AM} = \mu = E(x) = \int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x$

$$= \int_0^1 x \cdot 6x(1-x)\mathrm{d}x$$

$$= 6\int_0^1 (x^2-x^3)\mathrm{d}x$$

$$= 6\left|\frac{x^3}{3}-\frac{x^4}{4}\right|_0^1$$

$$= 6\left(\frac{1}{3}-\frac{1}{4}\right)$$

$$= \frac{1}{2}$$

(iii) $\frac{1}{H} = \int_{-\infty}^{\infty} \frac{1}{x} f(x)\mathrm{d}x$

$$= \int_0^1 \frac{1}{x}\cdot 6x(1-x)\mathrm{d}x$$

$$= 6\int_0^1 (1-x)\mathrm{d}x$$

$$= 6\left|x-\frac{x^2}{2}\right|_0^1$$

$$= 6\left(1-\frac{1}{2}\right)$$

$$= 3$$

$$H = \frac{1}{3}$$

(iv) $\int_0^M f(x)\,\mathrm{d}x = \frac{1}{2}$

$$\int_0^M 6x(1-x)\,\mathrm{d}x = \frac{1}{2}$$

$$6\int_0^M (x-x^2)\mathrm{d}x = \frac{1}{2}$$

$$6\left|\frac{x^2}{2}-\frac{x^3}{3}\right|_0^M = \frac{1}{2}$$

$$6\left(\frac{M^2}{2}-\frac{M^3}{3}\right)=\frac{1}{2}$$

$$3M^2-2M^3=\frac{1}{2}$$

$$6M^2-4M^3=1$$

$$4M^3-6M^2+1=0$$

$$M=\frac{1}{2},\frac{1}{2}\pm\frac{\sqrt{3}}{2}$$

The values $M=\frac{1}{2}\pm\frac{\sqrt{3}}{2}$ lie outside (0, 1).

Hence, $M=\frac{1}{2}$

(v) Mode is the value of x for which $f(x)$ is maximum. For $f(x)$ to be maximum, $f'(x)=0$ and $f''(x)<0$.

$$f'(x)=0$$

$$6-12x=0$$

$$x=\frac{1}{2}$$

$$f''(x)=-12<0$$

Hence, $f(x)$ is maximum at $x=\frac{1}{2}$

Mode $=\frac{1}{2}$

As the mean, median and mode are equal, the distribution is symmetrical.

(vi) $E(X^2)=\int_{-\infty}^{\infty} x^2+f(x)\,dx$

$$=\int_0^1 x^2\cdot 6x(1-x)\,dx$$

$$=6\int_0^1 (x^3-x^4)\,dx$$

$$=6\left|\frac{x^4}{4}-\frac{x^5}{5}\right|_0^1$$

$$=6\left(\frac{1}{4}-\frac{1}{5}\right)$$

$$=\frac{3}{10}$$

$$\text{Var}(X)=E(X^2)-\{E(X)\}^2$$

$$=\frac{3}{10}-\left(\frac{1}{2}\right)^2$$

$$=\frac{1}{20}$$

$$SD=\sqrt{\text{Var}(X)}=\sqrt{\frac{1}{20}}=\frac{1}{2\sqrt{5}}$$

(vii) Mean deviation about the mean

$$MD=\int_{-\infty}^{\infty}|x-\mu|\,f(x)\,dx$$

$$=\int_0^1\left|x-\frac{1}{2}\right|6x(1-x)\,dx$$

$$=\int_0^{\frac{1}{2}}\left(\frac{1}{2}-x\right)6x(1-x)\,dx+\int_{\frac{1}{2}}^{1}\left(x-\frac{1}{2}\right)6x(1-x)\,dx$$

$$=\int_0^{\frac{1}{2}}(3x-9x^2+6x^3)\,dx+\int_{\frac{1}{2}}^{1}(-3x+9x^2-6x^3)\,dx$$

$$=\left|\frac{3x^2}{2}-3x^3+\frac{3x^4}{2}\right|_0^{\frac{1}{2}}+\left|-\frac{3x^2}{2}+3x^3-\frac{3x^4}{2}\right|_{\frac{1}{2}}^{1}$$

$$=\left(\frac{3}{8}-\frac{3}{8}+\frac{3}{32}\right)+\left(-\frac{3}{2}+3-\frac{3}{2}\right)-\left(-\frac{3}{8}+\frac{3}{8}-\frac{3}{32}\right)$$

$$=\frac{3}{16}$$

Example 15

Prove that geometric mean G of the distribution

$$f(x)=6(2-x)(x-1),\ 1\le x\le 2$$

is given by $6\log(16G)=19$.

Solution

$$\log G = \int_{-\infty}^{\infty} (\log x) f(x)\,dx$$

$$= \int_{1}^{2} (\log x)\, 6(2-x)(x-1)\,dx$$

$$= -6\int_{1}^{2} (x^2 - 3x + 2)\log x\,dx$$

$$= -6\left[\left|\left(\frac{x^3}{3} - \frac{3x^2}{2} + 2x\right)\log x\right|_1^2 - \int_1^2 \left(\frac{x^3}{3} - \frac{3x^2}{2} + 2x\right)\frac{1}{x}\,dx\right]$$

$$= -6\left[\left(\frac{8}{3} - 6 + 4\right)\log 2 - \int_1^2 \left(\frac{x^2}{3} - \frac{3x}{2} + 2\right)dx\right]$$

$$= -6\left[\frac{2}{3}\log 2 - \left|\frac{x^3}{9} - \frac{3x^2}{4} + 2x\right|_1^2\right]$$

$$= -6\left[\frac{2}{3}\log 2 - \left(\frac{8}{9} - 3 + 4\right) + \left(\frac{1}{9} - \frac{3}{4} + 2\right)\right]$$

$$= -6\left[\frac{2}{3}\log 2 - \frac{17}{9} + \frac{49}{36}\right]$$

$$= -4\log 2 + \frac{19}{6}$$

$$\log G + 4\log 2 = \frac{19}{6}$$

$$\log(G \times 2^4) = \frac{19}{6}$$

$$\log(16G) = \frac{19}{6}$$

Example 16

The probability distribution of a random variable X is

$$f(x) = k\sin\frac{\pi}{5}x, \quad 0 \le x \le 5$$

Determine the constant k and obtain the median and quartiles of the distribution.

Solution

Since $f(x)$ is a probability distribution,

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$\int_0^5 k\sin\frac{\pi}{5}x\,dx = 1$$

$$k\left|\frac{-\cos\frac{\pi}{5}x}{\frac{\pi}{5}}\right|_0^5 = 1$$

$$\frac{5k}{\pi}(-\cos\pi + \cos 0) = 1$$

$$\frac{5k}{\pi}[-(-1) + 1] = 1$$

$$\frac{10k}{\pi} = 1$$

$$k = \frac{\pi}{10}$$

Hence,

$$f(x) = \frac{\pi}{10}\sin\frac{\pi}{5}x, \quad 0 \le x \le 5$$

The r^{th} quartile Q_r is given by

$$\int_{-\infty}^{Q_r} f(x)\,dx = \frac{r}{4}, \quad r = 1, 2, 3$$

$$\int_0^{Q_r} \frac{\pi}{10}\sin\frac{\pi}{5}x\,dx = \frac{r}{4}$$

$$\frac{\pi}{10}\left|\frac{-\cos\frac{\pi}{5}x}{\frac{\pi}{5}}\right|_0^{Q_r} = \frac{r}{4}$$

$$\frac{1}{2}\left(-\cos\frac{\pi}{5}Q_r + \cos 0\right) = \frac{r}{4}$$

$$-\cos\frac{\pi}{5}Q_r + 1 = \frac{r}{2}$$

$$\cos\frac{\pi}{5}Q_r = 1-\frac{r}{2}$$

$$\frac{\pi}{5}Q_r = \cos^{-1}\left(1-\frac{r}{2}\right)$$

$$Q_r = \frac{5}{\pi}\cos^{-1}\left(1-\frac{r}{2}\right)$$

$$Q_1 = \frac{5}{\pi}\cos^{-1}\left(1-\frac{1}{2}\right) = \frac{5}{\pi}\cos^{-1}\left(\frac{1}{2}\right) = \frac{5}{\pi}\left(\frac{\pi}{3}\right) = \frac{5}{3}$$

$$Q_2 = \frac{5}{\pi}\cos^{-1}(1-1) = \frac{5}{\pi}\cos^{-1}(0) = \frac{5}{\pi}\left(\frac{\pi}{2}\right) = \frac{5}{2}$$

$$Q_3 = \frac{5}{\pi}\cos^{-1}\left(1-\frac{3}{2}\right) = \frac{5}{\pi}\cos^{-1}\left(-\frac{1}{2}\right) = \frac{5}{\pi}\left(\frac{2\pi}{3}\right) = \frac{10}{3}$$

$$\text{Median} = Q_2 = \frac{5}{2}$$

Example 17

Find the median, mode and quartile deviation of continuous random variable X, given that its density functions is

$$f(x) = \frac{k}{1+x^2}, \quad -\infty < x < \infty.$$

Solution

(i) Since $f(x)$ is a probability density function,

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$$

$$\int_{-\infty}^{\infty} \frac{k}{1+x^2}\mathrm{d}x = 1$$

$$2k\int_{0}^{\infty} \frac{1}{1+x^2}\mathrm{d}x = 1 \quad \left[\because \int_{-a}^{a} f(x)dx = 2\int_{0}^{a} f(x)dx, \text{ if } f(x) \text{ is even function}\right]$$

$$2k\left|\tan^{-1}x\right|_{0}^{\infty} = 1$$

$$2k(\tan^{-1}\infty - \tan^{-1}0) = 1$$

$$2k\left(\frac{\pi}{2}\right) = 1$$

$$k = \frac{1}{\pi}$$

Hence, $$f(x) = \frac{1}{\pi(1+x^2)}, -\infty < x < \infty$$

(ii) The r^{th} quartile Q_r is given by

$$\int_{-\infty}^{Q_r} f(x)\mathrm{d}x = \frac{r}{4}, \quad r = 1, 2, 3$$

$$\int_{-\infty}^{Q_r} \frac{1}{\pi(1+x^2)}\mathrm{d}x = \frac{r}{4}$$

$$\frac{1}{\pi}\left|\tan^{-1}x\right|_{-\infty}^{Q_r} = \frac{r}{4}$$

$$\frac{1}{\pi}\left[\tan^{-1}Q_r - \tan^{-1}(-\infty)\right] = \frac{r}{4}$$

$$\frac{1}{\pi}\left[\tan^{-1}Q_r - \left(-\frac{\pi}{2}\right)\right] = \frac{r}{4}$$

$$\tan^{-1}Q_r = \frac{\pi}{4}(r-2)$$

$$Q_r = \tan\left\{\frac{\pi}{4}(r-2)\right\}$$

$$Q_1 = \tan\left(-\frac{\pi}{4}\right) = -1$$

$$Q_2 = \tan(0) = 0$$

$$Q_3 = \tan\left(\frac{\pi}{4}\right) = 1$$

$$QD = \frac{1}{2}(Q_3 - Q_1)$$

$$= \frac{1}{2}[1-(-1)]$$

$$= 1$$

(iii) Median $Q_2 = 0$

(iv) Mode is the value of x for which $f(x)$ is maximum. For $f(x)$ to be maximum $f'(x) = 0$ and $f''(x) < 0$.

$$f'(x) = 0$$

$$-\frac{2x}{\pi(1+x^2)^2} = 0$$

$$x = 0$$

$$f''(x) = -\frac{2}{\pi}\left[\frac{(1+x^2)^2 - x\cdot 2(1+x^2)\cdot 2x}{(1+x^2)^4}\right]$$

$$= \frac{2}{\pi}\left[\frac{3x^2-1}{(1+x^2)^3}\right]$$

$$f''(0) = -\frac{2}{\pi} < 0$$

Hence, $f(x)$ is maximum at $x = 0$.
Mode = 0

Example 18

Find the mean, variance and the coefficients β_1, β_2 *of the distribution* $f(x) = kx^2e^{-x}$, $0 < x < \infty$

Solution

Since $f(x)$ is a probability density function,

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$$

$$\int_{0}^{\infty} kx^2e^{-x}\mathrm{d}x = 1$$

$$k\left|x^2(-e^{-x}) - 2x\,e^{-x} + 2(-e^{-x})\right|_0^{\infty} = 1$$

$$k(2e^0) = 1$$

$$k = \frac{1}{2}$$

Hence, $f(x) = \frac{1}{2}x^2e^{-x}, 0 < x < \infty$

$$\mu_2' = \int_{-\infty}^{\infty} x^r f(x)\mathrm{d}x$$

$$= \int_0^{\infty} x^r \frac{1}{2}x^2e^{-x}\mathrm{d}x$$

$$= \frac{1}{2}\int_0^{\infty} e^{-x}x^{r+2}\mathrm{d}x$$

$$= \frac{1}{2}\left|\overline{r+3}\right.$$

$$= \frac{1}{2}(r+2)!$$

$$\mu_1' = \frac{1}{2}(3!) = 3$$

$$\mu_2' = \frac{1}{2}(4!) = 12$$

$$\mu_3' = \frac{1}{2}(5!) = 60$$

$$\mu_4' = \frac{1}{2}(6!) = 360$$

$$\mu_2 = \mu_2' - (\mu_1')^2 = 12 - (3)^2 = 3$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3 = 60 - 3(12)(3) + 2(3)^3 = 6$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$$

$$= 360 - 4(6)(3) + 6(12)(3)^2 - 3(3)^4$$

$$= 45$$

$$\text{Mean} = \mu_1' = 3$$

$$\text{Variance} = \mu_2 = 3$$

$$\beta_1 = \frac{\mu_3^2}{\mu_3^2} = \frac{(6)^2}{(3)^3} = \frac{4}{3}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{45}{(3)^2} = 5$$

Example 19

The probability density function of a random variable X is given by $f(x) = kx\,(2-x)$, $0 \le x \le 2$. *Find mean, variance* β_1 *and* β_2.

Solution

Since $f(x)$ is a probability density function,

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$$

$$\int_0^2 k\,x(2-x)\,dx = 1$$

$$k\int_0^2 (2x-x^2)\,dx = 1$$

$$k\left|x^2-\frac{x^3}{3}\right|_0^2 = 1$$

$$k\left(4-\frac{8}{3}\right) = 1$$

$$k = \frac{3}{4}$$

Hence, $f(x) = \frac{3}{4}x(2-x),\ 0 \le x \le 2$

$$\mu_r' = \int_{-\infty}^{\infty} x^r f(x)\,dx$$

$$= \int_0^2 x^r \frac{3}{4}x(2-x)\,dx$$

$$= \frac{3}{4}\int_0^2 x^{r+1}(2-x)\,dx$$

$$= \frac{2(2^{r+1})}{(r+2)(r+3)}$$

$$\mu_1' = \frac{3(2^2)}{(3)(4)} = 1$$

$$\mu_2' = \frac{3(2^3)}{(4)(5)} = \frac{6}{5}$$

$$\mu_3' = \frac{3(2^4)}{(5)(6)} = \frac{8}{5}$$

$$\mu_4' = \frac{3(2^5)}{(6)(7)} = \frac{16}{7}$$

$$\mu_2 = \mu_2' - (\mu_1')^2 = \frac{6}{5} - 1 = \frac{1}{5}$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3 = \frac{8}{5} - 3\left(\frac{6}{5}\right)(1) + 2 = 0$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$$

$$= \frac{16}{7} - 4\left(\frac{8}{5}\right)(1) + 6\left(\frac{6}{5}\right)(1)^2 - 3(1)^4$$

$$= \frac{3}{35}$$

$$\text{Mean} = \mu_1' = 1$$

$$\text{Variance} = \mu_2 = \frac{1}{5}$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\frac{3}{35}}{\left(\frac{1}{5}\right)^2} = \frac{15}{7}$$

Example 20

Show that for the symmetrical distribution

$$f(x) = \frac{2a}{\pi}\left(\frac{1}{a^2+x^2}\right), \quad -a \le x \le a$$

$$\mu_2 = \frac{a^2(4-\pi)}{\pi} \quad \text{and} \quad \mu_4 = a^4\left(1-\frac{8}{3\pi}\right)$$

Solution

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{-a}^{a} \frac{2a}{\pi}\left(\frac{1}{a^2+x^2}\right)dx$$

$$= \frac{2a}{\pi}\left|\frac{1}{a}\tan^{-1}\frac{x}{a}\right|_{-a}^{a}$$

$$= \frac{2}{\pi}\left|\tan^{-1}\frac{x}{a}\right|_{-a}^{a}$$

$$= \frac{2}{\pi}\left[\tan^{-1}(1) - \tan^{-1}(-1)\right]$$

$$= 1$$

Hence, $f(x)$ represents a probability density function.

$$\mu_1' = \int_{-\infty}^{\infty} x\, f(x)\,dx$$

$$= \int_{-a}^{a} x \frac{2a}{\pi}\left(\frac{1}{a^2+x^2}\right)dx$$

$$= \frac{2a}{\pi}\int_{-a}^{a} \frac{x}{a^2+x^2}\,dx$$

$$= \frac{2a}{\pi}\left|\frac{1}{2}\log(a^2+x^2)\right|_{-a}^{a}$$

$$= 0 \qquad [\because \text{integrand is an odd function of } x]$$

$$u_2' = \int_{-\infty}^{\infty} x^2\, f(x)\,dx$$

$$= \int_{-a}^{a} x^2 \frac{2a}{\pi}\left(\frac{1}{a^2+x^2}\right)dx$$

$$= \frac{2a}{\pi}\int_{-a}^{a} \frac{x^2}{a^2+x^2}\,dx$$

$$= \frac{4a}{\pi}\int_{0}^{a} \frac{x^2+a^2-a^2}{a^2+x^2}\,dx$$

$$= \frac{4a}{\pi}\int_{0}^{a}\left(1-\frac{a^2}{a^2+x^2}\right)dx$$

$$= \frac{4a}{\pi}\left|x - a\tan^{-1}\frac{x}{a}\right|_0^a$$

$$= \frac{4a}{\pi}(a - a\tan^{-1}1)$$

$$= \frac{4a}{\pi}\left(a - a\frac{\pi}{4}\right)$$

$$= \frac{a^2(4-\pi)}{\pi}$$

$$\mu_2 = \mu_2' - (\mu_1')^2 = \frac{a^2(4-\pi)}{\pi} - 0 = \frac{a^2(4-\pi)}{\pi}$$

$$\mu_4 = \mu_4' \qquad (\because \mu_1' = 0)$$

$$\mu_4 = \int_{-\infty}^{\infty} x^4 \cdot f(x)\,dx$$

$$= \int_{-a}^{a} x^4 \cdot \frac{2a}{\pi}\left(\frac{1}{a^2+x^2}\right)dx$$

$$= \frac{2a}{\pi}\int_{-a}^{a} \frac{x^4}{a^2+x^2}\,dx$$

$$= \frac{4a}{\pi}\int_{0}^{a}\left(x^2 - a^2 + \frac{a^4}{a^2+x^2}\right)dx$$

$$= \frac{4a}{\pi}\left|\frac{1}{3}x^3 - a^2x + a^3\tan^{-1}\frac{x}{a}\right|_0^a$$

$$= \frac{4a}{\pi}\left(\frac{a^3}{3} - a^3 + a^3\tan^{-1}1\right)$$

$$= \frac{4a}{\pi}\left(\frac{a^3}{3} - a^3 + a^3\frac{\pi}{4}\right)$$

$$= a^4\left(1 - \frac{8}{3\pi}\right)$$

EXERCISE 3.4

1. If the probability density function is given by

$f(x) = kx^2(1-x^3) \qquad 0 \le x \le 1$
$= 0 \qquad$ otherwise

Find (i) k, (ii) $P\left(0 < X < \frac{1}{2}\right)$, (iii) $\bar{X}$, and (iv) σ^2.

$\left[\textbf{Ans.: (i) } 6 \textbf{ (ii) } \frac{15}{64} \textbf{ (iii) } \frac{9}{14} \textbf{ (iv) } \frac{9}{245}\right]$

2. If the probability density function of a random variable is given by

$f(x) = kx \qquad 0 \le x \le 2$
$= 2k \qquad 2 \le x \le 4$
$= 6k - kx \qquad 4 \le x \le 6$

Find (i) k, (ii) $P(1 \le X \le 3)$, and (iii) $\bar{X}$.

$\left[\textbf{Ans.: (i) } \frac{1}{2} \textbf{ (ii) } \frac{1}{3} \textbf{ (iii) } \frac{383}{36}\right]$

3. If the probability density of a random variable is given by

$$f(x) = kxe^{-\frac{x}{3}} \quad x > 0$$
$$= 0 \quad x \le 0$$

Find (i) k, (ii) $\overline{X}$, and (iii) σ^2.

$\left[\textbf{Ans.:} \text{ (i) } \frac{1}{9} \text{ (ii) } 6 \text{ (iii) } 18\right]$

4. A continuous random variable has the probability density function

$$f(x) = 2e^{-2x} \quad x > 0$$
$$= 0 \quad x \le 0$$

Find (i) $E(X)$, (ii) $E(\overline{X})$, (iii) Var (X), and (iv) SD of X.

$\left[\textbf{Ans.:} \text{ (i) } \frac{1}{2} \text{ (ii) } \frac{1}{2} \text{ (iii) } \frac{1}{4} \text{ (iv) } \frac{1}{2}\right]$

5. A random variable X has the pdf

$$f(x) = \frac{k}{1+x^2}, \ -\infty < X < \infty$$

Determine (i) k, (ii) $P(X \ge 0)$, (iii) mean, and (iv) variance.

$\left[\textbf{Ans.:} \text{ (i) } \frac{1}{\pi} \text{ (ii) } \frac{1}{2} \text{ (iii) 0 (iv) does not exist}\right]$

6. The distribution function of a continuous random variable X is given by $F(x) = 1-(1+x)e^{-x}, x \ge 0$. Find (i) pdf, (ii) mean, and (iii) variance.

$\left[\textbf{Ans.:} \text{ (i) } f(x) = xe^{-x}, x \ge 0 \text{ (ii) 2 (iii) 2}\right]$

7. If $f(x)$ is the probability density function of a continuous random variable, find k, mean, and variance.

$$f(x) = kx^2 \quad 0 \le x \le 1$$
$$= (2-x)^2 \quad 1 \le x \le 2$$

$\left[\textbf{Ans.:} \ 2, \frac{11}{12}, 0.626\right]$

8. A continuous random variable X has the probability density function given by

$$f(x) = 2ax + b \quad 0 \le x \le 2$$
$$= 0 \quad \text{otherwise}$$

If the mean of the distribution is 3, find the constants a and b.

$\left[\textbf{Ans.:} \ \frac{3}{2}, -\frac{5}{2}\right]$

9. If X is a continuous random variable with probability density function given by

$$f(x) = k(x - x^3) \quad 0 \le x \le 1$$
$$= 0 \quad \text{otherwise}$$

Find (i) k, (ii) mean, (iii) variance, and (iv) median.

$\left[\textbf{Ans.:} \text{ (i) } \frac{1}{2} \text{ (ii) 0.06 (iii) 0.04 (iv) 2}\right]$

10. The probability density function of a random variable is given by

$$f(x) = 0 \quad x < 2$$
$$= \frac{2x+3}{18} \quad 2 \le x \le 4$$
$$= 0 \quad x > 4$$

Find the mean and variance.

$\left[\textbf{Ans.:} \text{ (i) } \frac{83}{27}, 0.33\right]$

11. A continuous random variable X has the probability density function

$$f(x) = x^3 \quad 0 \le x \le 1$$
$$= (2-x)^3 \quad 1 \le x \le 2$$
$$= 0 \quad \text{otherwise}$$

Find $P(0.5 \le X \le 1.5)$ and mean of the distribution.

$\left[\textbf{Ans.:} \ \frac{15}{32}, \frac{1}{2}\right]$

12. The probability density function of a continuous random variable X is given by

$$f(x) = kx(2-x) \quad 0 \le x \le 2$$

Find k, mean, and variance.

$\left[\textbf{Ans.:} \ \frac{3}{4}, 1, \frac{1}{5}\right]$

13. If the density function of a continuous random variable X is given by $f(x) = \lambda e^{-\lambda(x-a)}$, $a \le x < \infty$, show that $\beta_1 = 4$ and $\beta_2 = 9$.

14. If the continuous random variable has the density function $f(x)=\frac{kx}{(1+x)^3}, x \geq 0$, find the value of k, median and mode.

$\left[\text{Ans.: } 2, 1+\sqrt{2}, \frac{1}{2}\right]$

15. The density function of a continuous random variable X is given by $f(x)=\frac{3}{4}x(2-x), 0 \leq x \leq 2$. Find the mean, median, mode, harmonic mean, MD about mean and SD.

$\left[\text{Ans.: } 1, 1, 1, \frac{2}{3}, \frac{3}{8}, \frac{1}{\sqrt{5}}\right]$

16. The density function of a continuous random variable X is given by

$$f(x)\begin{cases} \frac{1}{16}(3+x^2) & -3 \leq x \leq -1 \\ \frac{1}{16}(6-2x^2) & -1 \leq x \leq 1 \\ \frac{1}{16}(3-x^2) & 1 \leq x \leq 3 \end{cases}$$

Find the mean, SD and MB about the mean.

$\left[\text{Ans.: } 0, 1, \frac{13}{16}\right]$

3.8 EXPECTED VALUES OF TWO DIMENSIONAL RANDOM VARIABLES

If (X, Y) is a two dimensional discrete random variable with joint probability mass function $P(x_i, y_j) = p_{ij}$, then the mathematical expectation of a function $g(x, y)$ is given by

$$E[g(X,Y)] = \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} g(x_i, y_j)p_{ij}$$
$$= \sum_{x}\sum_{y} g(x,y)f(x),y)$$

If (X, Y) is a two dimensional continuous random variable with joint probability density function $f(x, y)$, then the mathematical expectation of a function $g(x, y)$ is given by

$$E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)\,dx\,dy$$

3.8.1 Properties of Expected Values of Two Dimensional Random Variables

(i) If X and Y are random variables, then $E(X+Y) = E(X) + E(Y)$ provided all the expectations exist.

(ii) If X and Y are independent random variables then $E(XY) = E(X) \cdot E(Y)$.

3.8.2 Conditional Expectation and Conditional Variance

If (X,Y) is a two dimensional discrete random variable with joint probability mass function p_{ij} then the conditional expectations of $g(X, Y)$ is given by

$$E\{g(X,Y)/Y=y_j\} = \sum_{i=1}^{\infty} g(x_i, y_j)P(X=x_i/Y=y_j)$$
$$= \sum_{i=1}^{\infty} \frac{g(x_i,y_j)P(X=x_i, Y=y_j)}{P(Y=y_j)}$$
$$= \sum_{i=1}^{\infty} g(x_i,y_j)\frac{p_{ij}}{p_{*j}}$$

In particular, the conditional expectation of a discrete random variable X given $Y = y_j$ is given by

$$E(X/Y=y_j) = \sum_{i=1}^{\infty} x_i P(X=x_i/Y=y_j)$$

The conditional variance of X given $Y = y_j$ is given by

$$\mathrm{Var}(X/Y=y_j) = E\left[\{X-E(X/Y=y_j)\}^2/Y=y_j\right]$$

If (X, Y) is a two-dimensional continuous random variable with joint probability density function $f(x, y)$, then the conditional expectation of $g(X, Y)$ is given by

$$E\{g(X,Y)/Y=y\} = \int_{-\infty}^{\infty} g(x,y)f(x/y)\,dx$$
$$= \int_{-\infty}^{\infty} \frac{g(x,y)f(x,y)\,dx}{f_Y(y)}$$

In particular, the conditional expectation of X given $Y = y$ is given by

$$E(X/Y=y) = \frac{\int_{-\infty}^{\infty} xf(x,y)\,dx}{f_Y(y)}$$

Similarly, $$E(Y/X=x) = \frac{\int_{-\infty}^{\infty} yf(x,y)\,dx}{f_X(x)}$$

The conditional variance of X is given by

$$\text{Var}\,(X/Y = y) = E[\{X - E(X/Y = y)\}^2 / \, Y = y]$$

Similarly, $$\text{Var}(Y/X = x) = E[\{Y - E(Y/X = x)\}^2 \, /X = x]$$

3.8.3 Properties of Conditional Expectation

(i) If x and y are independent random variables, then

$$E(Y/X) = E(Y)$$

and $$E(X/Y) = E(X)$$

(ii) $E(XY) = E[X.\, E(Y/X)]$

(iii) $E(X^2Y^2) = E(X^2.\, E(Y^2/X)]$

Example 1

Given a pair of discrete random variable X and Y whose joint probability distribution is given by

X Y	2	4
1	0.1	0.15
2	0.2	0.3
3	0.1	0.15

Find the expected value of the function g(X, Y) *given that* g(X, Y) = 2X + Y.

Solution

$$\begin{aligned} E[g(x,y)] &= \sum_x \sum_y g(x,y) f(x,y) \\ &= \sum_x \sum_y (2x+y) f(x,y) \\ &= \{2(2)+1\}0.1 + \{2(4)+1\}0.15 \\ &\quad + \{2(2)+2\}0.2 + \{2(4)+2\}0.3 \\ &\quad + \{2(2)+3\}0.1 + \{2(4)+3\}0.15 \\ &= 8.4 \end{aligned}$$

Example 2

Let X and Y be two random variables each taking values −1, 0 and 1 and having the joint probability distribution as given below:

X Y	−1	0	1	Total p(y)
−1	0	0.1	0.1	0.2
0	0.2	0.2	0.2	0.6
1	0	0.1	0.1	0.2
Total p(x)	0.2	0.4	0.4	1.0

(i) Show that X and Y have different expectation.

(ii) Find E(XY)

(iii) Find Var(X) *and* Var(Y).

(iv) Given that Y = 0, what is the conditional probability distribution of X?

Solution

(i) $$\begin{aligned} E(X) &= \sum xp(x) \\ &= -1(0.2) + 0(0.4) + 1(0.4) \\ &= 0.2 \\ E(Y) &= \sum yp(y) \\ &= -1(0.2) + 0(0.6) + 1(0.2) \\ &= 0 \\ E(X) &\neq E(Y) \end{aligned}$$

(ii) $$\begin{aligned} E(XY) &= \sum x_i y_j p_{ij} \\ &= (-1)(-1)(0) + (0)(-1)(0.1) + (1)(-1)(0.1) \\ &\quad + (-1)(0)0.2 + (0)(0)(0.2) + (1)(0)(0.2) \\ &\quad + (-1)(1)(0) + (0)(1)(0.1) + (1)(1)(0.1) \\ &= 0 \end{aligned}$$

(iii) $$\begin{aligned} E(X^2) &= \sum x^2 p(x) \\ &= (-1)^2(0.2) + 0(0.4) + (1)^2(0.4) \\ &= 0.6 \\ \text{Var}(X) &= E(X^2) - \{E(X)\}^2 \\ &= 0.6 - (0.2)^2 \\ &= 0.56 \\ E(Y^2) &= \sum y^2 p(y) \\ &= (-1)^2(0.2) + 0(0.6) + (1)^2(0.2) \\ &= 0.4 \end{aligned}$$

$$\mathrm{Var}(Y) = E(Y^2) - \{E(Y)\}^2$$
$$= 0.4 - 0$$
$$= 0.4$$

(iv) $$P(X=-1/Y=0) = \frac{P(X=-1, Y=0)}{P(Y=0)}$$
$$= \frac{0.2}{0.6} = \frac{1}{3}$$
$$P(X=0/Y=0) = \frac{P(X=0, Y=0)}{P(Y=0)}$$
$$= \frac{0.2}{0.6} = \frac{1}{3}$$
$$P(X=1/Y=0) = \frac{P(X=1, Y=0)}{P(Y=0)}$$
$$= \frac{0.2}{0.6} = \frac{1}{3}$$

Example 3

If the joint pdf of (X,Y) is given by

$$f(x,y) = \begin{cases} \dfrac{16y}{x^3} & x > 2, 0 < y < 1 \\ 0 & \text{elsewhere} \end{cases}$$

then find E(X,Y).

Solution

$$E(X,Y) = \int_{-\infty}^{\infty} \cdot \int_{-\infty}^{\infty} xy f(x,y)\,dx\,dy$$
$$= \int_0^1 \int_2^{\infty} xy\left(\frac{16y}{x^3}\right) dx\,dy$$
$$= 16 \int_0^1 \int_2^{\infty} \left(\frac{y^2}{x^2}\right) dx\,dy$$
$$= 16 \int_0^1 y^2 \left| -\frac{1}{x} \right|_2^{\infty} dy$$
$$= 16 \int_0^1 \frac{1}{2} y^2 dy$$

$$= 8 \left| \frac{y^3}{3} \right|_0^1$$
$$= \frac{8}{3}(1-0)$$
$$= \frac{8}{3}$$

Example 4

The joint PDF of (X,Y) is given by

$$f(x, y) = 24xy, \quad x > 0, y > 0, x + y \le 1$$
$$= 0, \quad \text{elsewhere}$$

Find the conditional mean and variance of Y, given X.

Solution

The region of integration is ΔOAB.

In ΔOAB, along vertical strip PQ, limits of y: $y = 0$ to $y = 1 - x$ and x varies from $x = 0$ to $x = 1$.

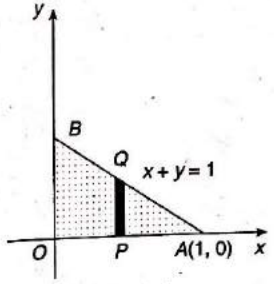


Fig. 3.4

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$$
$$= \int_0^{1-x} 24xy\,dy$$
$$= 24x \left| \frac{y^2}{2} \right|_0^{1-x}$$
$$= 12x(1-x)^2, \quad 0 \le x \le 1$$
$$f(y/x) = \frac{f(x,y)}{f_X(x)}$$
$$= \frac{24xy}{12x(1-x)^2}$$
$$= \frac{2y}{(1-x)^2}$$
$$E(Y/X=x) = \int_{-\infty}^{\infty} y f(y/x)\,dy$$
$$= \int_0^{1-x} y \cdot \frac{2y}{(1-x)^2} dy$$

$$=\frac{2}{(1-x)^2}\left|\frac{y^3}{3}\right|_0^{1-x}$$

$$=\frac{2}{3}(1-x)$$

$$E(Y^2/x)=\int_{-\infty}^{\infty} y^2 f(y/x)\mathrm{d}y$$

$$=\int_0^{1-x} y^2\frac{2y}{(1-x)^2}\mathrm{d}y$$

$$\mathrm{Var}(Y^2/x)=(E(Y^2/x)-\{E(Y/x)\}^2$$

$$=\frac{1}{2}(1-x)^2-\left\{\frac{2}{3}(1-x)\right\}^2$$

$$=\frac{1}{2}(1-x)^2-\frac{4}{9}(1-x)^2$$

$$=\frac{1}{18}(1-x)^2$$

$$=\frac{1}{96}\int_1^5 y^2\left|\frac{x^3}{3}\right|_0^4 \mathrm{d}y$$

$$=\frac{1}{288}\int_1^5 y^2\left|x^3\right|_0^4 \mathrm{d}y$$

$$=\frac{1}{288}\int_1^5 64y^2\mathrm{d}y$$

$$=\frac{2}{9}\left|\frac{y^3}{3}\right|_1^5$$

$$=\frac{2}{27}(125-1)$$

$$=\frac{248}{27}$$

(iv) $E(2X+3Y)=2E(X)+3E(Y)$

$$=2\left(\frac{8}{3}\right)+3\left(\frac{31}{9}\right)$$

$$=\frac{47}{3}$$

(v) $E(X^2)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^2 f(x,y)\mathrm{d}x\,\mathrm{d}y$

$$=\int_1^5\int_0^4 x^2\left(\frac{xy}{96}\right)\mathrm{d}x\,\mathrm{d}y$$

$$=\frac{1}{96}\int_1^5 y\left|\frac{x^4}{4}\right|_0^4 \mathrm{d}y$$

$$=\frac{1}{384}\int_1^5 256y\,\mathrm{d}y$$

$$=\frac{2}{3}\left|\frac{y^2}{2}\right|_1^5$$

$$=\frac{1}{3}(25-1)$$

$$=8$$

$$\mathrm{Var}(X)=E(X^2)-\{E(X)\}^2$$

$$=8-\left(\frac{8}{3}\right)^2$$

$$=\frac{8}{9}$$

(vi) $E(Y^2)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y^2 f(x,y)\mathrm{d}x\mathrm{d}y$

$$=\int_1^5\int_0^4 y^2\left(\frac{xy}{96}\right)\mathrm{d}x\,\mathrm{d}y$$

$$=\frac{1}{96}\int_1^5 y^3\left|\frac{x^2}{2}\right|_0^4 \mathrm{d}y$$

$$=\frac{1}{192}\int_1^5 16y^3\,\mathrm{d}y$$

$$=\frac{1}{12}\left|\frac{y^4}{4}\right|_1^5$$

$$=\frac{1}{48}(625-1)$$

$$=13$$

$$\begin{aligned}\text{Var}(Y) &= E(Y^2)-\{E(Y)\}^2\\ &=13-\left(\frac{31}{9}\right)^2\\ &=\frac{92}{81}\end{aligned}$$

(vii) $\text{Cov}(X,Y)=E(XY)-E(X)E(Y)$

$$\begin{aligned}&=\frac{248}{27}-\left(\frac{8}{3}\right)\left(\frac{31}{9}\right)\\ &=0\end{aligned}$$

Example 5

Two random variables X and Y have the following joint probability density function:

$$f(x,y)=\begin{cases}2-x-y, & 0\le x\le 1, 0\le y\le 1\\ 0 \quad , & \text{otherwise}\end{cases}$$

Find (i) *Marginal probability density function of X and Y.*

(ii) *Conditional density functions*

(iii) *Var(X)* and *Var(Y)*

(iv) *Covariance between X and Y*

Solution

(i)

$$\begin{aligned}f_X(x)&=\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y\\ &=\int_0^1 (2-x-y)\,\mathrm{d}y\\ &=\left|2y-xy-\frac{y^2}{2}\right|_0^1\\ &=\left(2-x-\frac{1}{2}\right)\\ &=\frac{3}{2}-x\end{aligned}$$

$$\therefore \quad f_X(x)=\begin{cases}\frac{3}{2}-x, & 0<x<1\\ 0 \quad , & \text{otherwise}\end{cases}$$

$$\text{Similarly, } f_Y(y)=\begin{cases}\frac{3}{2}-y, & 0<y<1\\ 0 \quad , & \text{otherwise}\end{cases}$$

(ii)

$$\begin{aligned}f_{X/Y}(x/y)&=\frac{f(x,y)}{f_Y(y)}\\ &=\frac{(2-x-y)}{\left(\frac{3}{2}-y\right)}, \qquad 0<(x,y)<1\end{aligned}$$

$$\begin{aligned}f_{Y/X}(y/x)&=\frac{f(x,y)}{f_X(x)}\\ &=\frac{(2-x-y)}{\left(\frac{3}{2}-x\right)}\end{aligned}$$

(iii)

$$\begin{aligned}E(X)&=\int_{-\infty}^{\infty} x\,f_X(x)\,\mathrm{d}x\\ &=\int_0^1 x\left(\frac{3}{2}-x\right)\mathrm{d}x\\ &=\left|\frac{3x^2}{4}-\frac{x^3}{3}\right|_0^1\\ &=\frac{3}{4}-\frac{1}{3}\\ &=\frac{5}{12}\end{aligned}$$

$$\begin{aligned}E(Y)&=\int_{-\infty}^{\infty} y\,f_Y(y)\,\mathrm{d}y\\ &=\int_0^1 y\left(\frac{3}{2}-y\right)\mathrm{d}y\\ &=\left|\frac{3y^2}{4}-\frac{y^3}{3}\right|_0^1\\ &=\frac{3}{4}-\frac{1}{3}\\ &=\frac{5}{12}\end{aligned}$$

$$E(X^2)=\int_{-\infty}^{\infty} x^2 f_X(x)\mathrm{d}x$$

$$=\int_0^1 x^2\left(\frac{3}{2}-x\right)\mathrm{d}x$$

$$=\left|\frac{x^3}{2}-\frac{x^4}{4}\right|_0^1$$

$$=\frac{1}{2}-\frac{1}{4}$$

$$=\frac{1}{4}$$

$$\mathrm{Var}(X)=E(X^2)=\{E(X)\}^2$$

$$=\frac{1}{4}-\left(\frac{5}{12}\right)^2$$

$$=\frac{11}{144}$$

Similarly, $\mathrm{Var}(Y)=\dfrac{11}{144}$

Example 6

If the joint pdf of (X, Y) is given by

$$f(x, y) = 24y(1-x),\ 0 \le y \le x \le 1,$$

then find E(XY).

Solution

The region of integration is ΔOAB. In ΔOAB, along horizontal strip $P'Q'$, Limits of x: $x = y$ to $x = 1$ and y varies from $y = 0$ to $y = 1$.

$$E(XY)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f(x, y)\,\mathrm{d}x\,\mathrm{d}y$$

$$=\int_0^1\int_y^1 xy\cdot 24y(1-x)\,\mathrm{d}x\,\mathrm{d}y$$

$$=24\int_0^1\int_y^1 xy^2(1-x)\,\mathrm{d}x\,\mathrm{d}y$$

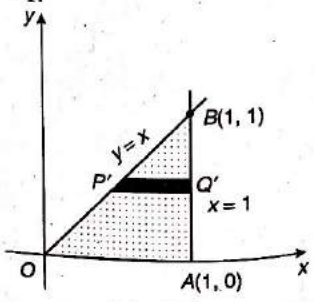


Fig. 3.5

$$=24\int_0^1 y^2\left|\frac{x^2}{2}-\frac{x^3}{3}\right|_y^1 \mathrm{d}y$$

$$=24\int_0^1 y^2\left(\frac{1}{2}-\frac{1}{3}-\frac{y^2}{2}+\frac{y^3}{3}\right)\mathrm{d}y$$

$$=24\int_0^1 y^2\left(\frac{1}{6}-\frac{y^2}{2}+\frac{y^3}{3}\right)\mathrm{d}y$$

$$=24\int_0^1\left(\frac{y^2}{6}-\frac{y^4}{2}+\frac{y^5}{3}\right)\mathrm{d}y$$

$$=24\left|\frac{y^3}{18}-\frac{y^5}{10}+\frac{46}{18}\right|_0^1$$

$$=14\left(\frac{1}{18}-\frac{1}{10}+\frac{1}{18}\right)$$

$$=\frac{4}{15}$$

Example 7

Two random variables have joint pdf

$$f(x,y)=\begin{cases}\dfrac{xy}{96}, & 0<x<4,\quad 1<y<5\\ 0\ , & \text{elsewhere}\end{cases}$$

Find (i) $E(X)$ (ii) $E(Y)$ (iii) $E(XY)$ (iv) $E(2X+3Y)$ (v) $\mathrm{Var}(X)$ (vi) $\mathrm{Var}(Y)$ (vii) $\mathrm{Cov}(X,Y)$

Solution

(i) $E(X)=\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\, f(x,y)\,\mathrm{d}x\,\mathrm{d}y$

$$=\int_1^5\int_0^4 x\left(\frac{xy}{96}\right)\mathrm{d}x\,\mathrm{d}y$$

$$=\frac{1}{96}\int_1^5 y\left|\frac{x^3}{3}\right|_0^4 \mathrm{d}y$$

$$=\frac{1}{96}\int_1^5 y\left(\frac{64}{3}\right)\mathrm{d}y$$

$$= \frac{2}{9}\left|\frac{y^2}{2}\right|_1^5$$

$$= \frac{2}{9}\left(\frac{25}{2} - \frac{1}{2}\right)$$

$$= \frac{8}{3}$$

(ii) $E(Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f(x,y)\,dx\,dy$

$$= \int_1^5\int_0^4 y\left(\frac{xy}{96}\right)dx\,dy$$

$$= \frac{1}{96}\int_1^5 y^2\left|\frac{x^2}{2}\right|_0^4 dy$$

$$= \frac{1}{96}\int_1^5 8y^2\,dy$$

$$= \frac{1}{12}\left|\frac{y^3}{3}\right|_1^5$$

$$= \frac{1}{36}(125-1)$$

$$= \frac{31}{9}$$

(iii) $E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f(x,y)\,dx\,dy$

$$= \int_1^5\int_0^4 xy\left(\frac{xy}{96}\right)dx\,dy$$

$$= \frac{1}{96}\int_1^5\int_0^4 x^2y^2\,dx\,dy$$

(iv) $E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f(x,y)\,dx\,dy$

$$= \int_0^1\int_0^1 xy(2-x-y)\,dx\,dy$$

$$= \int_0^1 y\left|x^2 - \frac{x^3}{3} - \frac{x^2y}{2}\right|_0^1 dy$$

$$= \int_0^1 y\left(1 - \frac{1}{3} - \frac{y}{2}\right)dy$$

$$= \int_0^1 \left(\frac{2y}{3} - \frac{y^2}{2}\right)dy$$

$$= \left|\frac{y^2}{3} - \frac{y^3}{6}\right|_0^1$$

$$= \frac{1}{3} - \frac{1}{6}$$

$$= \frac{1}{6}$$

(v) $\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)$

$$= \frac{1}{6} - \left(\frac{5}{12}\right)\left(\frac{5}{12}\right)$$

$$= -\frac{1}{144}$$

Example 8

Let $f(x, y) = 8xy, \quad 0 < x < y < 1$

$= 0 \quad , \quad$ elsewhere

Find (i) $E\,(Y/X = x)$ *(ii)* $E(XY/X = x)$ (iii) $\mathrm{Var}(Y/X = x)$.

Solution

The region of integration is ΔOAB. In ΔOAB, along vertical strip PQ, limits of y: $y = x$ to $y = 1$ and x varies from $x = 0$ to $x = 1$.

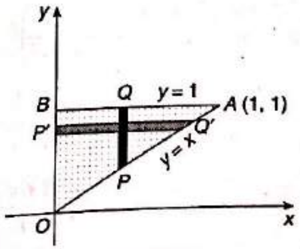


Fig. 3.6

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$$

$$= \int_x^1 8xy\,dy$$

$$= 8x\left|\frac{y^2}{2}\right|_x^1$$

$$= 4x(1-x^2) \quad 0 < x < 1$$

In ΔOAB, along horizontal strip $P'Q'$, Limits of x : $x = 0$ to $x = y$ and y varies from $y = 0$ to $y = 1$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$$

$$= \int_0^y 8xy\,dx$$

$$= 8y\left|\frac{x^2}{2}\right|_0^y$$

$$= 4y^3, \qquad 0 < y < 1$$

$$f_{X/Y}(x/y) = \frac{f(x,y)}{f_Y(y)}$$

$$= \frac{8xy}{4y^3}$$

$$= \frac{2x}{y^2}$$

$$f_{Y/X}(y/x) = \frac{f(x,y)}{f_X(x)}$$

$$= \frac{8xy}{4x(1-x^2)}$$

$$= \frac{2y}{1-x^2}$$

(i) $E(Y/X = x) = \int_{-\infty}^{\infty} y\, f_{Y/X}(y/x)\,dy$

$$= \int_x^1 y\left(\frac{2y}{1-x^2}\right)dy$$

$$= \frac{2}{1-x^2}\left|\frac{y^3}{3}\right|_x^1$$

$$= \frac{2}{3}\left(\frac{1-x^3}{1-x^2}\right)$$

$$= \frac{2}{3}\left(\frac{1+x+x^2}{1+x}\right)$$

(ii) $E(XY/X = x) = x\,E(Y/X = x)$

$$= \frac{2}{3}\frac{x(1+x+x^2)}{(1+x)}$$

(iii) $E(Y^2/X = x) = \int_{-\infty}^{\infty} y^2 f_{Y/X}(y/x)\,dy$

$$= \int_x^1 y^2\left(\frac{2y}{1-x^2}\right)dy$$

$$= \frac{2}{1-x^2}\left|\frac{y^4}{4}\right|_x^1$$

$$= \frac{1}{2}\left(\frac{1-x^4}{1-x^2}\right)$$

$$\text{Var}(Y/X = x) = E(Y^2/X = x) - \{E(Y/X = x)\}^2$$

$$= \frac{1+x^2}{2} - \left[\frac{2}{3}\left(\frac{1+x+x^2}{1+x}\right)\right]^2$$

$$= \frac{1+x^2}{2} - \frac{4}{9}\frac{(1+x+x^2)^2}{(1+x)^2}$$

EXERCISE 3.5

1. If the pdf of (X, Y) is given by

$$f(x, y) = 2 - x - y,\ 0 \le x \le y \le 1$$

Find $E(X)$ and $E(Y)$. $\left[\text{Ans.: } \frac{5}{12}, \frac{5}{12}\right]$

2. If $f(x,y) = \begin{cases} \frac{1}{\pi}, & 0 < x^2 + y^2 < 1 \\ 0, & x^2 + y^2 > 1 \end{cases}$ [Ans.: 0]

Find the covariance of X, Y.

3. Joint pdf of X and Y is given by

$$f(x, y) = 3(x + y) \qquad 0 \le x \le 1,\ 0 \le y \le 1$$

Find $E(Y/X = x)$ and Cov(X,Y). $\left[\text{Ans.: } \frac{(1-x)(x+2)}{3(1+x)}, -\frac{13}{320}\right]$

4. Let $f_{XY}(x, y) = e^{-(x+y)}$ $\qquad 0 \le x < \infty,\ 0 < y < \infty$

Find Cov(X, Y). [Ans.: 0]

5. If the joint pdf of (*X*, *Y*) is given by

$$f(x, y) = 2, \quad 0 \le x < y \le 1,$$

find the conditional mean and conditional variance of *X* given that $Y = y$.

$$\left[\textbf{Ans.: } \frac{y}{2}, \frac{y^2}{12}\right]$$

6. If the joint pdf of (*X*, *Y*) is given by

$$f(x, y) = 21\, x^2 y^3, \; 0 \le x < y \le 1$$

find the conditional mean and conditional variance of *X*, given that $Y = y$, $0 < y < 1$.

$$\left[\textbf{Ans.: } \frac{3y}{4}, \frac{3y^2}{80}\right]$$

7. If the joint pdf of (*X*, *Y*) is given by

$$f(x, y) = 3xy\,(x + y), \; 0 < x \le y \le 1,$$

verify that $E\{E(Y/X)\} = E(Y) = \dfrac{17}{24}$.

3.9 BOUNDS ON PROBABILITIES

If the probability distribution of a random variable is known *E*(*X*) and Var(*X*) can be computed. Conversely, if *E*(*X*) and Var(*X*) are known, probability distribution of *X* cannot be constructed and quantities such as $P\{|X - E(X)| \le k\}$ can not be evaluated. Several approximation techniques have been developed to yield upper and /or lower bounds to such probabilities. The most important of such techniques is Chebyshev's inequality.

3.10 CHEBYSHEV'S INEQUALITY

If *X* is a random variable with mean μ and variance σ^2, then for any positive umber *k*,

$$P\{|X - \mu| \ge k\sigma\} \le \frac{1}{k^2}$$

or

$$P\{|X - \mu| < k\sigma\} \ge 1 - \frac{1}{k^2}$$

Proof

Let *X* be a continuous random variable.

$$\sigma^2 = E[X - E(X)]^2$$

$$= E[X - \mu]^2 \qquad [\because \mu = E(X)]$$

$$= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\mathrm{d}x \quad \text{where } f(x) \text{ is pdf of } X.$$

$$= \int_{-\infty}^{\mu-k\sigma} (x-\mu)^2 f(x)\mathrm{d}x + \int_{\mu-k\sigma}^{\mu+k\sigma} (x-\mu)^2 f(x)\mathrm{d}x + \int_{\mu+k\sigma}^{\infty} (x-\mu)^2 f(x)\mathrm{d}x$$

$$\ge \int_{-\infty}^{\mu-k\sigma} (x-\mu)^2 f(x)\mathrm{d}x + \int_{\mu+k\sigma}^{\infty} (x-\mu)^2 f(x)\mathrm{d}x \qquad \text{...(1)}$$

We know that $x \le \mu - k\sigma$ and $x \ge \mu + k\sigma$

$\therefore \quad |x - \mu| \ge k\sigma$

Substituting in Eq. (1),

$$\sigma^2 \ge \int_{-\infty}^{\mu-k\sigma} k^2\sigma^2 f(x)\mathrm{d}x + \int_{\mu+k\sigma}^{\infty} k^2\sigma^2 f(x)\mathrm{d}x$$

$$= k^2\sigma^2 \left[\int_{-\infty}^{\mu-k\sigma} f(x)\mathrm{d}x + \int_{\mu+k\sigma}^{\infty} f(x)\mathrm{d}x\right]$$

$$= k^2\sigma^2 [P(X \le \mu - k\sigma) + P(X \ge \mu + k\sigma)]$$

$$= k^2\sigma^2 [P(X - \mu \le -k\sigma) + P(X - \mu \ge k\sigma)]$$

$$= k^2\sigma^2 P\{|X - \mu| \ge k\sigma\}$$

$$P\{|X - \mu| \ge k\sigma\} \le \frac{1}{k^2}$$

$$\because P\{|X - \mu| \ge k\sigma\} + P\{|X - \mu| < k\sigma\} = 1$$

$$P\{|X - \mu| < k\sigma\} = 1 - P\{|X - \mu| \ge k\sigma\}$$

$$\ge 1 - \frac{1}{k^2}$$

Note

1. If $k\sigma = c > 0$

$$P\{|X - \mu| \ge c\} \le \frac{\sigma^2}{c^2}$$

and

$$P\{|X - \mu| < c\} \ge 1 - \frac{\sigma^2}{c^2}$$

2. To find the lower bound of probabilities following form of Chebyshev's inequality is used:

$$P\{|X - \mu| < k\sigma\} \ge 1 - \frac{1}{k^2}$$

or $$P\{|X-\mu|<c\}\ge 1-\frac{\sigma^2}{c^2}$$

3. To find the upper bound of probabilities following form of Chebyshev's inequality is used;

$$P\{|X-\mu|\ge k\sigma\}\le\frac{1}{k^2}$$

or $$P\{|X-\mu|\ge c\}\le\frac{\sigma^2}{c^2}$$

Example 1

A random variable X has a mean $\mu = 12$ and a variance $\sigma^2 = 9$ and unknown probability distribution. Find $P(6 < X < 18)$.

Solution

$$\mu = 12,\quad \sigma^2 = 9$$
$$\sigma = 3$$

By Chebyshev's inequality,

$$P\{|X-\mu|<k\sigma\}\ge 1-\frac{1}{k^2}$$
$$P\{-k\sigma<X-\mu<k\sigma\}\ge 1-\frac{1}{k^2}$$
$$P\{\mu-k\sigma<X<\mu+k\sigma\}\ge 1-\frac{1}{k^2}$$
$$P\{12-3k<X<12+3k\}\ge 1-\frac{1}{k^2}$$

Comparing with $P(6 < X < 18)$,

$$12-3k=6$$
$$12+3k=18$$
$$\therefore\quad k=2$$
$$P\{6<X<18\}\ge 1-\frac{1}{4}$$
$$P\{6<X<18\}\ge\frac{3}{4}$$

Example 2

A random variable X has a mean 10 and a variance 4 and unknown probability distribution. Find the value of c such that $P\{|X-10|\ge c\}\le 0.04$.

Solution

$$\mu = 10,\ \sigma^2 = 4$$
$$\sigma = 2$$

By Chebyshev's inequality,

$$P\{|X-\mu|\ge k\sigma\}\le\frac{1}{k^2}$$

Comparing with $P\{|X-10|\ge c\}\le 0.04$,

$$\frac{1}{k^2}=0.04$$
$$k=5$$

and $$k\sigma=c$$
$$c=5(2)=10$$

Example 3

A random variable X has pdf $f(x)=e^{-x}, x\ge 0$. Use Chebyshev's inequality to show that $P\{|X-1|>2\}\le\frac{1}{4}$ and also, show that the actual probability is given by e^{-3}.

Solution

$$f(x)=e^{-x}$$

The random variable X follows exponential distribution with parameter $\lambda = 1$.

$$E(X)=\mu=\frac{1}{\lambda}=1$$
$$\text{Var}(X)=\sigma^2=\frac{1}{\lambda^2}=1$$

By Chebyshev's inequality,

$$P\{|X-\mu|>k\sigma\}\le\frac{1}{k^2}$$

Comparing with $P\{|X-\mu|>2\}$,

$$k\sigma = 2$$
$$k(1) = 2$$
$$k = 2$$

$\therefore$ $$P\{|X-1|>2\} \le \frac{1}{4}$$

The actual probability is given by

$$\begin{aligned} P\{|X-1|>2\} &= 1-P\{|X-1|\le 2\} \\ &= 1-P\{-1<X\le 3\} \\ &= 1-P\{0<X\le 3\} \\ &= 1-\int_0^3 e^{-x}\mathrm{d}x \\ &= 1-\left|e^{-x}\right|_0^3 \\ &= 1-e^{-3} \end{aligned}$$

Example 4

A random variable X is exponentially distributed with parameter 1. Use Chebyshev's inequality to show that $P\{-1\le X\le 3\}\ge\frac{3}{4}$. Find the actual probability also.

Solution

For an exponential distribution with parameter $\lambda = 1$,

$$E(X) = \mu = \frac{1}{\lambda} = 1$$
$$\text{Var}(X) = \sigma^2 = \frac{1}{\lambda^2} = 1$$
$$\sigma = 1$$

By Chebyshev's inequality,

$$P\{|X-\mu|<k\sigma\} \ge 1-\frac{1}{k^2}$$
$$P\{-k\sigma < X-\mu < k\sigma\} \ge 1-\frac{1}{k^2}$$

$$P\{\mu-k\sigma < X < \mu+k\sigma\} \ge 1-\frac{1}{k^2}$$
$$P\{1-k < X < 1+k\} \ge 1-\frac{1}{k^2}$$

Comparing with $P\{-1\le X\le 3\}\ge\frac{3}{4}$,

$$1-k=-1$$
$$k=2$$

$\therefore$ $$\begin{aligned} P\{-1\le X\le 3\} &\ge 1-\frac{1}{4} \\ &\ge \frac{3}{4} \end{aligned}$$

The actual probability is given by

$$\begin{aligned} P\{-1\le X\le 3\} &= P\{0\le X\le 3\} \quad [\because x>0 \text{ for exponential distribution}] \\ &= \int_0^3 f(x)\mathrm{d}x \\ &= \int_0^3 e^{-x}\mathrm{d}x \\ &= \left|-e^{-x}\right|_0^3 \\ &= -e^{-3}+e^{0} \\ &= 1-e^{-3} \\ &= 0.9502 \end{aligned}$$

Example 5

A fair dice is tossed 120 times. Use Chebyshev's inequality to find a lower bound for the probability of getting 80 to 120 sixes.

Solution

Let X be the random variable which denotes number of sixes obtained when a fair dice is tossed by 720 times.

$$n = 720$$

Probability of getting 6 in single toss

$$p = \frac{1}{6}$$
$$q = 1-p = 1-\frac{1}{6} = \frac{5}{6}$$

X follows a binomial distribution.

$$\mu = np-(720)\left(\frac{1}{6}\right)=120$$

$$\sigma^2 = npq = (720)\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)=100$$

$$\sigma = 10$$

By Chebyshev's inequality,

$$P\{|X-\mu|< k\sigma\} \ge 1-\frac{1}{k^2}$$

$$P\{-k\sigma < X-\mu < k\sigma\} \ge 1-\frac{1}{k^2}$$

$$P\{\mu-k\sigma < X < \mu+k\sigma\} \ge 1-\frac{1}{k^2}$$

$$P\{120-10k < X < 120+10k\} \ge 1-\frac{1}{k^2}$$

Comparing with $P\{80 < X < 120\}$,

$$120-10k = 80$$

$$k = 4$$

$$P\{80 < X < 120\} \ge 1-\frac{1}{4^2}$$

$$P\{80 < X < 120\} \ge \frac{15}{16}$$

Hence, the lower bound for probability $= \frac{15}{16}$

Example 6

Two dice are thrown once. If X is the sum of the numbers sharing up, prove that $P\{|X-7| \ge 3\} \le \frac{35}{34}$. *Compare this value with the exact probability.*

Solution

Let X_1 and X_2 be the random variables which denote the outcomes of first and second dice.

$$E(X_1) = E(X_2) = \frac{1}{6}(1+2+3+4+5+6) = \frac{7}{2}$$

$$E(X) = E(X_1)+E(X_2) = \mu = \frac{7}{2}+\frac{7}{2} = 7$$

$$E(X_1^2) = E(X_2^2) = \frac{1}{6}(1^2+2^2+3^2+4^2+5^2+6^2) = \frac{91}{6}$$

$$\text{Var}(X_1) = \text{Var}(X_2) = \frac{91}{6}-\left(\frac{7}{2}\right)^2 = \frac{35}{12}$$

$$\text{Var}(X) = \text{Var}(X_1+X_2) = (1)^2\,\text{Var}(X_1)+(1)^2\,\text{Var}(X_2)$$

$$\sigma^2 = \frac{35}{12}+\frac{35}{12} = \frac{35}{6}$$

$$\sigma = \sqrt{\frac{35}{6}}$$

By Chebyshev's inequality,

$$P\{|X-\mu| \ge k\sigma\} \le \frac{1}{k^2}$$

Comparing with $P\{|X-7| \ge 3\}$,

$$\mu = 7$$

$$k\sigma = 3$$

$$k\sqrt{\frac{35}{6}} = 3$$

$$k = 3\sqrt{\frac{6}{35}}$$

$$\therefore \quad P\{|X-7| \ge 3\} \le \frac{1}{\left(3\sqrt{\frac{6}{35}}\right)^2}$$

$$\le \frac{35}{54}$$

Actual probability is given by

$$P\{|X-7| \ge 3\} = P\{X = 1,2,3,4,10,11,12\}$$

$$= \frac{1}{36}+\frac{2}{36}+\frac{3}{36}+\frac{4}{36}+\frac{3}{36}+\frac{2}{36}+\frac{1}{36}$$

$$= \frac{4}{9}$$

Example 7

Use Chebyshev's inequality to find how many times a fair coin must be tossed in order that probability that the ratio of the number of heads *to the number of tosses will the between* 0.45 *and* 0.55 *will be at least* 0.95.

Solution

Let X be the random variable which denotes the number of heads obtained when a fair coin is tossed n times.

$$p = q = \frac{1}{2}$$

X follows a binomial distribution.

$$\text{Mean} = np \quad \text{and} \quad \text{Var}(X) = npq$$

Mean of required ratio $\frac{x}{n} = E\left(\frac{1}{n}X\right) = \frac{1}{n}E(X)$

$$= \frac{1}{n}np = p = \frac{1}{2}$$

$$\therefore \quad \mu = \frac{1}{2}$$

$$\text{Var}\left(\frac{X}{n}\right) = \left(\frac{1}{n}\right)^2 \text{Var}(X) = \frac{1}{n^2}npq = \frac{pq}{n}$$

$$\sigma = \sqrt{\frac{pq}{n}} = \sqrt{\frac{\frac{1}{2}\cdot\frac{1}{2}}{n}} = \frac{1}{2\sqrt{n}}$$

By Chebyshev's inequality,

$$P\left\{\left|\frac{X}{n} - \mu\right| < k\sigma\right\} \ge 1 - \frac{1}{k^2}$$

$$P\left\{-k\sigma < \frac{X}{n} - \mu < k\sigma\right\} \ge 1 - \frac{1}{k^2}$$

$$P\left\{\mu - k\sigma < \frac{X}{n} < \mu + k\sigma\right\} \ge 1 - \frac{1}{k^2}$$

But $P\left\{0.45 < \frac{X}{n} < 0.55\right\} \ge 0.95$

$$1 - \frac{1}{k^2} = 0.95$$

$$\frac{1}{k^2} = 0.05$$

$$k = \sqrt{20}$$

$$\mu - k\sigma = 0.45$$

$$0.5 - \left(\frac{1}{2\sqrt{n}}\right) = 0.45$$

$$n = 2000$$

Hence, the fair coin must be tossed 2000 times.

Example 8

If X is the number on a dice when it is thrown, prove that $P\{|X - \mu| \ge 2.5\} \le 0.47$, *where μ is the mean.*

Solution

Let x be the random variable which denotes the number on a dice. The probability function is

X	1	2	3	4	5	6
$P(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

$$E(X) = \mu = \sum xp(x)$$

$$= 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right)$$

$$= \frac{7}{2}$$

$$\text{Var}(X) = \sigma^2 = \sum x^2 p(x) - \mu^2$$

$$= 1\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 9\left(\frac{1}{6}\right) + 16\left(\frac{1}{6}\right) + 25\left(\frac{1}{6}\right) + 36\left(\frac{1}{6}\right) - \left(\frac{7}{2}\right)^2$$

$$= 2.9167$$

$$\sigma = 1.707$$

By Chebyshev's inequality,

$$P\{|X - \mu| > k\sigma\} < \frac{1}{k^2}$$

Comparing with $P\{|X - \mu| > 2.5\}$,

$$k\sigma = 2.5$$

$$k(1.707) = 2.5$$

$$k = 1.46$$

$$\therefore\ P\left\{|X-\mu|>2.5\right\}<\frac{1}{(1.46)^2}$$

$$P\left\{\{X-\mu\}>2.5\right\}<0.47$$

Example 9

The number of planes landing at an airport in a 30 minutes interval obeys the Poisson law with mean 25. Use Chebyshev's inequality to find the least chance that the number of planes landing within a given 30 minutes interval will be between 15 and 25.

Solution

Let x be a random variable which denotes the number of planes landing at an airport.

For Poisson distribution,

$$E(X)=\mu=25$$

$$\text{Var}(X)=\sigma^2=\mu=25$$

$$\sigma=5$$

By Chebyshev's inequality,

$$P\left\{|X-\mu|<k\sigma\right\}\geq 1-\frac{1}{k^2}$$

$$P\left\{-k\sigma<X-\mu<k\sigma\right\}\geq 1-\frac{1}{k^2}$$

$$P\left\{\mu-k\sigma<X<\mu+k\sigma\right\}\geq 1-\frac{1}{k^2}$$

$$P\left\{25-5k<X<25+5k\right\}\geq 1-\frac{1}{k^2}$$

Comparing with $P\{15<X<25\}$,

$$25-5k=15 \text{ and } 25+5k=25$$

$$k=2$$

$$\therefore P\{15<X<25\}\geq 1-\frac{1}{(2)^2}$$

$$\geq\frac{3}{4}$$

EXERCISE 3.6

1. A discrete random variable takes the values –1, 0, 1 with probability $\frac{1}{8},\frac{3}{4},\frac{1}{8}$ respectively. Find $P\left\{|X-\mu|\geq 25\right\}$.

$\left[\text{Ans.: }\frac{1}{4}\right]$

2. Use Chebyshev's inequality to prove that $P\{X=\mu\}=1$ if $\text{Var}(X)=0$.
3. If X is a random variable with $E(X)=3$ and $E(X^2)=13$, find the lower bound for $P(-2<X<8)$ using Chebyshev's inequality.

$\left[\text{Ans.: }\frac{21}{25}\right]$

4. Can we find a random variable for which $P\{\mu-2\sigma<X<\mu+2\sigma\}=0.6$?

[Ans.: No]

5. If X denotes the sum of the numbers obtained when 2 dice are drawn, obtain an upper bound for $P\{|X-7|\geq 4\}$. Compare with actual probability.

$\left[\text{Ans.: }\frac{35}{96},\frac{1}{6}\right]$

6. A fair dice is tossed 720 times. Use Chebyshev's inequality to find a lower bound for getting 100 to 140 sixes.

$\left[\text{Ans.: }\frac{3}{4}\right]$

7. A pair of dice is rolled 900 times and X denotes the number of times a total of 9 occurs. Find $P(80\leq X\leq 120)$ using Chebyshev's inequality.

$\left[\text{Ans.: }\frac{2}{9}\right]$

8. A discrete random variable X can assume the values x = 1, 2, 3, ... with probability 2^{-x}. Show that $P\{|X-2|\geq 2\}\leq\frac{1}{2}$, while the actual probability is $\frac{1}{8}$.
9. A random variable X has the *pmf* $P(X=1)=\frac{1}{18}$, $P(X=2)=\frac{16}{18}$, $P(X=3)=\frac{1}{18}$. Show that there is a value of c such that $P\left\{|X-\mu|\geq c\right\}=\frac{\sigma^2}{c^2}$,

so that, in general, the bound given by Chebyshev's inequality can not be improved.

10. Using Chebyshev's inequality find how many times a fair coin must be tossed in order that the probability of the ratio of number of heads to the number of tosses will lie between 0.4 and 0.6 will be at least 0.9.

[Ans.: 250]

11. Suppose that number of articles produced in a factory during a week is a random variable with mean 500 and variance 100. What can be said about the probability that a week's production will lie between 400 and 600.

[Ans.: At least 0.99]

CHAPTER 4

Correlation and Regression

Chapter Outline

4.1 Introduction
4.2 Correlation
4.3 Types of Correlations
4.4 Methods of Studying Correlation
4.5 Scatter Diagram
4.6 Simple Graph
4.7 Karl Pearson's Coefficient of Correlation
4.8 Properties of Coefficient of Correlation
4.9 Rank Correlation
4.10 Regression
4.11 Types of Regression
4.12 Methods of Studying Regression
4.13 Lines of Regression
4.14 Regression Coefficients
4.15 Properties of Regression Coefficients
4.16 Properties of Lines of Regression (Linear Regression)

4.1 INTRODUCTION

Correlation and regression are the most commonly used techniques for investigating the relationship between two quantitative variables. *Correlation* refers to the relationship of two or more variables. It measures the closeness of the relationship between the variables. *Regression* establishes a functional relationship between the variables. In correlation, both the variables x and y are random variables, whereas in regression, x is a random variable and y is a fixed variable. The coefficient of correlation is a relative measure whereas the regression coefficient is an absolute figure.

4.2 CORRELATION

Correlation is the relationship that exists between two or more variables. Two variables are said to be correlated if a change in one variable affects a change in the other variable. Such a data connecting two variables is called *bivariate data*. Thus, correlation is a statistical analysis which measures and analyses the degree or extent to which two variables fluctuate with reference to each other. Some examples of such a relationship are as follows:

1. Relationship between heights and weights.
2. Relationship between price and demand of commodity.
3. Relationship between rainfall and yield of crops.
4. Relationship between age of husband and age of wife.

4.3 TYPES OF CORRELATIONS

Correlation is classified into four types:

1. Positive and negative correlations
2. Simple and multiple correlations
3. Partial and total correlations
4. Linear and nonlinear correlations

4.3.1 Positive and Negative Correlations

Depending on the variation in the variables, correlation may be positive or negative.

1. Positive Correlation If both the variables vary in the same direction, the correlation is said to be positive. In other words, if the value of one variable increases, the value of the other variable also increases, or, if value of one variable decreases, the value of the other variable decreases, e.g., the correlation between heights and weights of group of persons is a positive correlation.

Height (cm)	150	152	155	160	162	165
Weight (kg)	60	62	64	65	67	69

2. Negative Correlation If both the variables vary in the opposite direction, correlation is said to be negative. In other words, if the value of one variable increases, the value of the other variable decreases, or, if the value of one variable decreases, the value of the other variable increases, e.g., the correlation between the price and demand of a commodity is a negative correlation.

Price (₹ per unit)	10	8	6	5	4	1
Demand (units)	100	200	300	400	500	600

4.3.2 Simple and Multiple Correlations

Depending upon the study of the number of variables, correlation may be simple or multiple.

1. Simple Correlation When only two variables are studied, the relationship is described as simple correlation, e.g., the quantity of money and price level, demand and price, etc.

2. Multiple Correlation When more than two variables are studied, the relationship is described as multiple correlation, e.g., relationship of price, demand, and supply of a commodity.

4.3.3 Partial and Total Correlations

Multiple correlation may be either partial or total.

1. Partial Correlation When more than two variables are studied excluding some other variables, the relationship is termed as partial correlation.

2. Total Correlation When more than two variables are studied without excluding any variables, the relationship is termed total correlation.

4.3.4 Linear and Nonlinear Correlations

Depending upon the ratio of change between two variables, the correlation may be linear or nonlinear.

1. Linear Correlation If the ratio of change between two variables is constant, the correlation is said to be linear. If such variables are plotted on a graph paper, a straight line is obtained, e.g.,

Milk (*l*)	5	10	15	20	25	30
Curg (kg)	2	4	6	8	10	12

2. Nonlinear Correlation If the ratio of change between two variables is not constant, the correlation is said to nonlinear. The graph of a nonlinear or curvilinear relationship will be a curve, e.g.,

Advertising expenses (₹ in lacs)	3	6	9	12	15
Sales (₹ in lacs)	10	12	15	15	16

4.4 METHODS OF STUDYING CORRELATION

There are two different methods of studying correlation, (1) Graphic methods (2) Mathematical methods.

Graphic methods are (a) scatter diagram, and (b) simple graph.

Mathematical methods are (a) Karl Pearson's coefficient of correlation, and (b) Spearman's rank coefficient of correlation.

4.5 SCATTER DIAGRAM

The scatter diagram is a diagrammatic representation of bivariate data to find the correlation between two variables. There are various correlationships between two variables represented by the following scatter diagrams.

1. Perfect Positive Correlation If all the plotted points lie on a straight line rising from the lower left-hand corner to the upper right-hand corner, the correlation is said to be perfectly positive (Fig. 4.1).

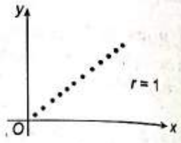


Fig. 4.1

2. Perfect Negative Correlation If all the plotted points lie on a straight line falling from the upper-left hand corner to the lower right-hand corner, the correlation is said to be perfectly negative (Fig. 4.2).

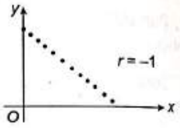


Fig. 4.2

3. High Degree of Positive Correlation If all the plotted points lie in the narrow strip, rising from the lower left-hand corner to the upper right-hand corner, it indicates a high degree of positive correlation (Fig. 4.3).

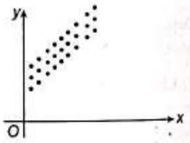


Fig. 4.3

4. High Degree of Negative Correlation If all the plotted points lie in a narrow strip, falling from the upper left-hand corner to the lower right-hand corner, it indicates the existence of a high degree of negative correlation (Fig. 4.4).

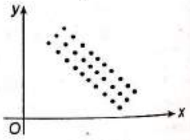


Fig. 4.4

5. No Correlation If all the plotted points lie on a straight line parallel to the *x*-axis or *y*-axis or in a haphazard manner, it indicates the absence of any relationship between the variables (Fig. 4.5).

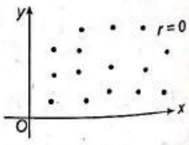


Fig. 4.5

Merits of a Scatter Diagram

1. It is simple and nonmathematical method to find out the correlation between the variables.

2. It gives an indication of the degree of linear correlation between the variables.
3. It is easy to understand.
4. It is not influenced by the size of extreme items.

4.6 SIMPLE GRAPH

A *simple graph* is a diagrammatic representation of bivariate data to find the correlation between two variables. The values of the two variables are plotted on a graph paper. Two curves are obtained, one for the variable *x* and the other for the variable *y*. If both the curves move in the same direction, the correlation is said to be positive. If both the curves move in the opposite direction, the correlation is said to be negative. This method is used in the case of a time series. It does not reveal the extent to which the variables are related.

4.7 KARL PEARSON'S COEFFICIENT OF CORRELATION

The coefficient of correlation is the measure of correlation between two random variables X and Y, and is denoted by r.

$$r = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

where cov (X, Y) is the covariance of variables X and Y,

σ_X is the standard deviation of variable X,

and σ_Y is the standard deviation of variable Y.

This expression is known as Karl Pearson's coefficient of correlation or Karl Pearson's product-moment coefficient of correlation.

$$\text{cov}(X,Y) = \frac{1}{n}\sum (x-\bar{x})(y-\bar{y})$$

$$\sigma_X = \sqrt{\frac{\sum (x-\bar{x})^2}{n}}$$

$$\sigma_Y = \sqrt{\frac{\sum (y-\bar{y})^2}{n}}$$

$$\therefore \quad r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2}\sqrt{\sum (y-\bar{y})^2}}$$

The above expression can be further modified.

Expanding the terms,

$$r=\frac{\sum(xy-x\bar{y}-\bar{x}y+\bar{x}\,\bar{y})}{\sqrt{\sum\left(x^2-2x\bar{x}+\bar{x}^2\right)}\sqrt{\sum\left(y^2-2y\bar{y}+\bar{y}^2\right)}}$$

$$=\frac{\sum xy-\bar{y}\sum x-\bar{x}\sum y+\bar{x}\,\bar{y}\sum 1}{\sqrt{\sum x^2-2\bar{x}\sum x+\bar{x}^2\sum 1}\sqrt{\sum y^2-2\bar{y}\sum y+\bar{y}^2\sum 1}}$$

$$=\frac{\sum xy-\frac{\sum y}{n}\sum x-\frac{\sum x}{n}\sum y+\frac{\sum x}{n}\frac{\sum y}{n}\cdot n}{\sqrt{\sum x^2-2\frac{\sum x}{n}\sum x+\left(\frac{\sum x}{n}\right)^2 n}\sqrt{\sum y^2-2\frac{\sum y}{n}\sum y+\left(\frac{\sum y}{n}\right)^2 n}}$$

$$=\frac{\sum xy-\frac{\sum x\sum y}{n}}{\sqrt{\sum x^2-\frac{\left(\sum x\right)^2}{n}}\sqrt{\sum y^2-\frac{\left(\sum y\right)^2}{n}}}$$

4.8 PROPERTIES OF COEFFICIENT OF CORRELATION

1. The coefficient of correlation lies between −1 and 1, i.e., $-1 \le r \le 1$.

Proof Let $\bar{x}$ and $\bar{y}$ be the mean of x and y series and σ_x and σ_y be their respective standard deviations.

$$\text{Let } \sum\left(\frac{x-\bar{x}}{\sigma_x}\pm\frac{y-\bar{y}}{\sigma_y}\right)^2\ge 0 \quad \left[\because \text{ sum of squares of real quantities cannot be negative}\right]$$

$$\frac{\sum(x-\bar{x})^2}{\sigma_x^2}+\frac{\sum(y-\bar{y})^2}{\sigma_y^2}\pm\frac{2\sum(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y}\ge 0$$

$$n+n\pm 2nr\ge 0$$

$$2n\pm 2nr\ge 0$$

$$2n(1\pm r)\ge 0$$

$$1\pm r\ge 0$$

$$\text{i.e.,}\quad 1+r\ge 0 \quad\text{or}\quad 1-r\ge 0$$

$$r\ge -1 \quad\text{or}\quad r\le 1$$

Hence, the coefficient of correlation lies between −1 and 1, i.e., $-1 \le r \le 1$.

2. Correlation coefficient is independent of change of origin and change of scale.

Proof Let $d_x=\frac{x-a}{h}$, $d_y=\frac{y-b}{k}$

$$x=a+hd_x,\quad y=b+kd_y$$

where a, b, h (>0) and k(>0) are constants.

$$x=a+hd_x\Rightarrow\bar{x}=a+h\bar{d}_x\Rightarrow x-\bar{x}=h(d_x-\bar{d}_x)$$

$$y=b+kd_y\Rightarrow\bar{y}=b+h\,\bar{d}_y\Rightarrow y-\bar{y}=k(d_y-\bar{d}_y)$$

$$r_{xy}=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}$$

$$=\frac{\sum h(d_x-\bar{d}_x)\,k(d_y-\bar{d}_y)}{\sqrt{\sum h^2(d_x-\bar{d}_x)^2}\sqrt{\sum k^2(d_y-\bar{d}_y)^2}}$$

$$=\frac{\sum(d_x-\bar{d}_x)(d_y-\bar{d}_y)}{\sqrt{\sum(d_x-\bar{d}_x)^2}\sqrt{(d_y-\bar{d}_y)^2}}$$

$$=r_{d_x d_y}$$

Hence, the correlation coefficient is independent of change of origin and change of scale.

Note Since correlation coefficient is independent of change of origin and change of scale,

$$r=\frac{\sum d_x d_y-\frac{\sum d_x\sum d_y}{n}}{\sqrt{\sum d_x^2-\frac{\left(\sum d_x\right)^2}{n}}\sqrt{\sum d_y^2-\frac{\left(\sum d_y\right)^2}{n}}}$$

3. Two independent variables are uncorrelated.

Proof If random variables X and Y are independent,

$$\sum(x-\bar{x})(y-\bar{y})=0 \text{ or } \operatorname{cov}(X,Y)=0$$

$$\therefore\qquad r=0$$

Thus, if X and Y are independent variables, they are uncorrelated.

Note The converse of the above property is not true, i.e., two uncorrelated variables may not be independent.

Example 1

Calculate the correlation coefficient between x and y using the following data:

x	2	4	5	6	8	11
y	18	12	10	8	7	5

Solution

$n = 6$

x	y	x^2	y^2	xy
2	18	4	324	36
4	12	16	144	48
5	10	25	100	50
6	8	36	64	48
8	7	64	49	56
11	5	121	25	55
$\sum x = 36$	$\sum y = 60$	$\sum x^2 = 266$	$\sum y^2 = 706$	$\sum xy = 293$

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}\sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}}$$

$$= \frac{293 - \frac{(36)(60)}{6}}{\sqrt{266 - \frac{(36)^2}{6}}\sqrt{706 - \frac{(60)^2}{6}}}$$

$$= -0.9203$$

Note $\sum x$, $\sum y$, $\sum x^2$, $\sum y^2$, $\sum xy$ can be directly obtained with the help of scientific calculator.

Example 2

Calculate the coefficient of correlation from the following data:

x	12	9	8	10	11	13	7
y	14	8	6	9	11	12	3

Solution

$n = 7$

x	y	x^2	y^2	xy
12	14	144	196	168
9	8	81	64	72
8	6	64	36	48
10	9	100	81	90
11	11	121	121	121
13	12	169	144	156
7	3	49	9	21
$\sum x = 70$	$\sum y = 63$	$\sum x^2 = 728$	$\sum y^2 = 651$	$\sum xy = 676$

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}\sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}}$$

$$= \frac{676 - \frac{(70)(63)}{7}}{\sqrt{728 - \frac{(70)^2}{7}}\sqrt{651 - \frac{(63)^2}{7}}}$$

$$= 0.949$$

Example 3

Calculate the coefficient of correlation for the following data:

x	9	8	7	6	5	4	3	2	1
y	15	16	14	13	11	12	10	8	9

Solution

$n = 9$

x	y	x^2	y^2	xy
9	15	81	225	135
8	16	64	256	128
7	14	49	196	98
6	13	36	169	78
5	11	25	121	55
4	12	16	144	48
3	10	9	100	30
2	8	4	64	16
1	9	1	81	9
$\sum x = 45$	$\sum y = 108$	$\sum x^2 = 285$	$\sum y^2 = 1356$	$\sum xy = 597$

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}\sqrt{\sum y^2 - \frac{\left(\sum y\right)^2}{n}}}$$

$$= \frac{597 - \frac{(45)(108)}{9}}{\sqrt{285 - \frac{(45)^2}{9}}\sqrt{1356 - \frac{(108)^2}{9}}}$$

$$= 0.95$$

Example 4

Calculate the correlation coefficient between the following data:

x	5	9	13	17	21
y	12	20	25	33	35

Solution

$n = 5$

$$\bar{x} = \frac{\sum x}{n} = \frac{65}{5} = 13$$

$$\bar{y} = \frac{\sum y}{n} = \frac{125}{5} = 25$$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
5	12	−8	−13	64	169	104
9	20	−4	−5	16	25	20
13	25	0	0	0	0	0
17	33	4	8	16	64	32
21	35	8	10	64	100	80
$\sum x = 65$	$\sum y = 125$	$\sum(x-\bar{x}) = 0$	$\sum(y-\bar{y}) = 0$	$\sum(x-\bar{x})^2 = 160$	$\sum(y-\bar{y})^2 = 358$	$\sum(x-\bar{x})(y-\bar{y}) = 236$

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}$$

$$= \frac{236}{\sqrt{160}\sqrt{358}}$$

$$= 0.986$$

Note Since $\sum x$, $\sum y$, $\sum x^2$, $\sum y^2$, $\sum xy$ can be directly obtained with the help of scientific calculator, correlation coefficient can be calculated without using mean.

Example 5

Calculate the correlation coefficient between for the following values of demand and the corresponding price of a commodity:

Demand in Quintals	65	66	67	67	68	69	70	72
Price in rupees per kg	67	68	65	68	72	72	69	71

Solution

Let the demand in quintal be denoted by x and the price in rupees per kg be denoted by y.

$$n = 8$$

$$\bar{x} = \frac{\sum x}{n} = \frac{544}{8} = 68$$

$$\bar{y} = \frac{\sum y}{n} = \frac{552}{8} = 69$$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
65	67	−3	−2	9	4	6
66	68	−2	−1	4	1	2
67	65	−1	−4	1	16	4
67	68	−1	−1	1	1	1
68	72	0	3	0	9	0
69	72	1	3	1	9	3
70	69	2	0	4	0	0
72	71	4	2	16	4	8
$\Sigma x = 544$	$\Sigma y = 552$	$\Sigma(x-\bar{x}) = 0$	$\Sigma(y-\bar{y}) = 0$	$\Sigma(x-\bar{x})^2 = 36$	$\Sigma(y-\bar{y})^2 = 44$	$\Sigma(x-\bar{x})(y-\bar{y}) = 24$

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}$$

$$= \frac{24}{\sqrt{36}\sqrt{44}}$$

$$= 0.603$$

Example 6

Calculate the coefficient of correlation for the following pairs of x and y:

x	17	19	21	26	20	28	26	27
y	23	27	25	26	27	25	30	33

Solution

Let $a = 23$ and $b = 27$ be the assumed means of x and y series respectively.

$$d_x = x - a = x - 23$$

$$d_y = y - b = y - 27$$

$$n = 8$$

x	y	d_x	d_y	d_x^2	d_y^2	$d_x d_y$
17	23	−6	−4	36	16	24
19	27	−4	0	16	0	0
21	25	−2	−2	4	4	4
26	26	3	−1	9	1	−3
20	27	−3	0	9	0	0
28	25	5	−2	25	4	−10
26	30	3	3	9	9	9
27	33	4	6	16	36	24
		$\Sigma d_x = 0$	$\Sigma d_y = 0$	$\Sigma d_x^2 = 124$	$\Sigma d_y^2 = 70$	$\Sigma d_x d_y = 48$

$$r = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sqrt{\sum d_x^2 - \frac{\left(\sum dx\right)^2}{n}}\sqrt{\sum d_y^2 - \frac{\left(\sum d_y\right)^2}{n}}}$$

$$= \frac{48-0}{\sqrt{124-0}\sqrt{70-0}}$$

$$= 0.515$$

Note Since Σx, Σy, Σx^2, Σy^2, Σxy can be directly obtained with the help of scientific calculator, the correlation coefficient can be calculated without using assumed mean.

Example 7

Calculate the correlation coefficient from the following data:

x	23	27	28	29	30	31	33	35	36	39
y	18	22	23	24	25	26	28	29	30	32

Solution

Let $a = 30$ and $b = 25$ be the assumed means of x and y series respectively.

$$d_x = x - a = x - 30$$

$$d_y = y - b = x - 25$$

$$n = 10$$

x	y	d_x	d_y	d_x^2	d_y^2	$d_x d_y$
23	18	–7	–7	49	49	49
27	22	–3	–3	9	9	9
28	23	–2	–2	4	4	4
29	24	–1	–1	1	1	1
30	25	0	0	0	0	0
31	26	1	1	1	1	1
33	28	3	3	9	9	9
35	29	5	4	25	16	20
36	30	6	5	36	25	30
39	32	9	7	81	49	63
		$\sum d_x = 11$	$\sum d_y = 7$	$\sum d_x^2 = 215$	$\sum d_y^2 = 163$	$\sum d_x d_y = 186$

$$r = \frac{\sum d_x d_y - \dfrac{\sum d_x \sum d_y}{n}}{\sqrt{\sum d_x^2 - \dfrac{\left(\sum d_x\right)^2}{n}}\sqrt{\sum d_y^2 - \dfrac{\left(\sum d_y\right)^2}{n}}}$$

$$= \frac{186 - \dfrac{(11)(7)}{10}}{\sqrt{215 - \dfrac{(11)^2}{10}}\sqrt{163 - \dfrac{(7)^2}{10}}}$$

$$= 0.996$$

Example 8

Calculate the coefficient of correlation between the ages of cars and annual maintenance costs.

Age of cars (year)	2	4	6	7	8	10	12
Annual maintenance cost (₹)	1600	1500	1800	1900	1700	2100	2000

Solution

Let the ages of cars in years be denoted by x and annual maintenance costs in rupees be denoted by y.

Let $a = 7$ and $b = 1800$ be the assumed means of x and y series respectively.

Let $h = 1$, $k = 100$

$$d_x = \frac{x - a}{h} = \frac{x - 7}{1} = x - 7$$

$$d_y = \frac{y - b}{k} = \frac{y - 1800}{100}$$

$$n = 7$$

x	y	d_x	d_y	d_x^2	d_y^2	$d_x d_y$
2	1600	–5	–2	25	4	10
4	1500	–3	3	9	9	9
6	1800	–1	0	1	0	0
7	1900	0	1	0	1	0
8	1700	1	–1	1	1	–1
10	2100	3	3	9	9	9
12	2000	5	2	25	4	10
		$\sum d_x = 0$	$\sum d_y = 0$	$\sum d_x^2 = 70$	$\sum d_y^2 = 28$	$\sum d_x d_y = 37$

$$r = \frac{\sum d_x d_y - \dfrac{\sum d_x \sum d_y}{n}}{\sqrt{\sum d_x^2 - \dfrac{\left(\sum d_x\right)^2}{n}}\sqrt{\sum d_y^2 - \dfrac{\left(\sum d_y\right)^2}{n}}}$$

$$= \frac{37 - 0}{\sqrt{70 - 0}\sqrt{28 - 0}}$$

$$= 0.836$$

Example 9

Calculate Karl Pearson's coefficient of correlation for the data given below:

x	10	14	18	22	26	30
y	18	12	24	6	30	36

Solution

Let $a = 22$ and $b = 24$ be the assumed means of x and y series respectively.

Let $h = 4$, $k = 6$

$$d_x = \frac{x-a}{h} = \frac{x-22}{4}$$

$$d_y = \frac{y-b}{k} = \frac{y-24}{6}$$

$$n = 6$$

x	y	d_x	d_y	d_x^2	d_y^2	$d_x d_y$
10	18	−3	−1	9	1	3
14	12	−2	−2	4	4	4
18	24	−1	0	1	0	0
22	6	0	−3	0	9	0
26	30	1	1	1	1	1
30	36	2	2	4	4	4
		$\Sigma d_x = -3$	$\Sigma d_y = -3$	$\Sigma d_x^2 = 19$	$\Sigma d_y^2 = 19$	$\Sigma d_x d_y = 12$

$$r = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sqrt{\sum d_x^2 - \frac{\left(\sum d_x\right)^2}{n}} \sqrt{\sum d_y^2 - \frac{\left(\sum d_y\right)^2}{n}}}$$

$$= \frac{12 - \frac{(-3)(-3)}{6}}{\sqrt{19 - \frac{(-3)^2}{6}} \sqrt{19 - \frac{(-3)^2}{6}}}$$

$$= 0.6$$

Example 10

The coefficient of correlation between two variables X and Y is 0.48. The covariance is 36. The variance of X is 16. Find the standard deviation of Y.

Solution

$r = 0.48$, $\quad \text{cov}(X, Y) = 36$, $\quad \sigma_X^2 = 16$

$\therefore \quad \sigma_X = 4$

$$r = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

$$0.48 = \frac{36}{4\sigma_Y}$$

$$\therefore \quad \sigma_Y = 18.75$$

Example 11

Given $n = 10$, $\sigma_X = 5.4$, $\sigma_Y = 6.2$, and sum of the product of deviations from the mean of x and y is 66. Find the correlation coefficient.

Solution

$$n = 10, \ \sigma_X = 5.4, \ \sigma_Y = 6.2$$

$$\sum (x-\bar{x})(y-\bar{y}) = 66$$

$$\sigma_X = \sqrt{\frac{\sum (x-\bar{x})^2}{n}}$$

$$5.4 = \sqrt{\frac{\sum (x-\bar{x})^2}{10}}$$

$$\therefore \quad \sum (x-\bar{x})^2 = 291.6$$

$$\sigma_Y = \sqrt{\frac{\sum (y-\bar{y})^2}{n}}$$

$$6.2 = \sqrt{\frac{\sum (y-\bar{y})^2}{10}}$$

$$\therefore \quad \sum (y-\bar{y})^2 = 384.4$$

$$r = \frac{\sqrt{\sum (x-\bar{x})(y-\bar{y})}}{\sqrt{\sum (x-\bar{x})^2} \sqrt{\sum (y-\bar{y})^2}}$$

$$= \frac{66}{\sqrt{291.6} \sqrt{384.4}}$$

$$= 0.197$$

Example 12

From the following information, calculate the value of n.

$$\sum x = 4, \sum y = 4, \sum x^2 = 44, \sum y^2 = 44, \sum xy = -40, r = -1$$

Solution

$$r=\frac{\sum xy-\frac{\sum x\sum y}{n}}{\sqrt{\sum x^2-\frac{\left(\sum x\right)^2}{n}}\sqrt{\sum y^2-\frac{\left(\sum y\right)^2}{n}}}$$

$$-1=\frac{-40-\frac{(4)(4)}{n}}{\sqrt{44-\frac{(4)^2}{n}}\sqrt{44-\frac{(4)^2}{n}}}$$

$$\therefore \quad n=8$$

Example 13

From the following data, find the number of items n.

$$r=0.5, \sum(x-\overline{x})(y-\overline{y})=120, \sigma_Y=8, \sum(x-\overline{x})^2=90$$

Solution

$$\sigma_Y=\sqrt{\frac{\sum(y-\overline{y})^2}{n}}$$

$$8=\sqrt{\frac{\sum(y-\overline{y})^2}{n}}$$

$$\sum(y-\overline{y})^2=64\,n$$

$$r=\frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2}\sqrt{\sum(y-\overline{y})^2}}$$

$$0.5=\frac{120}{\sqrt{90}\sqrt{64n}}$$

$$\therefore \quad n=10$$

Example 14

Calculate the correlation coefficient between x and y from the following data:

$$n=10, \sum x=140, \sum y=150, \sum(x-10)^2=180$$

$$\sum(y-15)^2=215, \sum(x-10)(y-15)=60$$

Solution

$$\sum d_x^2=\sum(x-10)^2=180$$

$$\sum d_y^2=\sum(y-15)^2=215$$

$$\sum d_x\,d_y=\sum(x-10)(y-15)=60$$

$$a=10$$

$$b=15$$

$$n=10$$

$$\overline{x}=\frac{\sum x}{n}=\frac{140}{10}=14$$

$$\overline{y}=\frac{\sum y}{n}=\frac{150}{10}=15$$

$$\overline{x}=a+\frac{\sum d_x}{n}$$

$$14=10+\frac{\sum d_x}{10}$$

$$\therefore \quad \sum d_x=40$$

$$\overline{y}=b+\frac{\sum d_y}{n}$$

$$15=15+\frac{\sum d_y}{10}$$

$$\therefore \quad \sum d_y=0$$

$$r=\frac{\sum d_x d_y-\frac{\sum d_x\sum d_y}{n}}{\sqrt{\sum d_x^2-\frac{\left(\sum d_x\right)^2}{n}}\sqrt{\sum d_y^2-\frac{\left(\sum d_y\right)^2}{n}}}$$

$$=\frac{60-\frac{(40)(0)}{10}}{\sqrt{180-\frac{(40)^2}{10}}\sqrt{215-\frac{0}{10}}}$$

$$=0.915$$

Example 15

A computer operator while calculating the coefficient between two variates x and y for 25 pairs of observations obtained the following constants:

$$n = 25, \sum x = 125, \sum x^2 = 650, \sum y = 100,$$
$$\sum y^2 = 460, \sum xy = 508$$

It was later discovered at the time of checking that he had copied down two pairs as (6, 14) *and* (8, 6) *while the correct pairs were* (8,12) *and* (6, 8). *Obtain the correct value of the correlation coefficient.*

Solution

$$n = 25$$

$$\text{Corrected} \sum x = \text{Incorrect} \sum x - (\text{Sum of incorrect } x) + (\text{Sum of correct } x)$$
$$= 125 - (6+8) + (8+6)$$
$$= 125$$

Similarly,

$$\text{Corrected} \sum y = 100 - (14+6) + (12+8) = 100$$
$$\text{Corrected} \sum x^2 = 650 - (6^2+8^2) + (8^2+6^2) = 650$$
$$\text{Corrected} \sum y^2 = 460 - (14^2+6^2) + (12^2+8^2) = 436$$
$$\text{Corrected} \sum xy = 508 - (84+48) + (96+48) = 520$$

Correct value of correlation coefficient

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}\sqrt{\sum y^2 - \frac{\left(\sum y\right)^2}{n}}}$$
$$= \frac{520 - \frac{(125)(100)}{25}}{\sqrt{650 - \frac{(125)^2}{25}}\sqrt{436 - \frac{(100)^2}{25}}}$$
$$= 0.67$$

EXERCISE 4.1

1. Draw a scatter diagram to represent the following data:

x	2	4	5	6	8	11
y	18	12	10	8	7	5

Calculate the coefficient of correlation between x and y.

[Ans.: –0.92]

2. Find the coefficient of correlation between x and y for the following data:

x	10	12	18	24	23	27
y	13	18	12	25	30	10

[Ans.: 0.223]

3. From the following information relating to the stock exchange quotations for two shares *A* and *B*, ascertain by using Pearson's coefficient of correlation how shares *A* and *B* are correlated in their prices?

Price share (*A*) ₹	160	164	172	182	166	170	178
Price share (*B*) ₹	292	280	260	234	266	254	230

[Ans.: –0.96]

4. Find the correlation coefficient between the income and expenditure of a wage earner.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul
Income	46	54	56	56	58	60	62
Expenditure	36	40	44	54	42	58	54

[Ans.: 0.769]

5. From the following data, examine whether the input of oil and output of electricity can be said to be correlated.

Input of oil	6.9	8.2	7.8	4.8	9.6	8.0	7.7
Output of Electricity	1.9	3.5	6.5	1.3	5.5	3.5	2.2

[Ans.: 0.696]

6. For the following data, show that cov $(x, x^2) = 0$.

x	–3	–2	–1	0	1	2	3
x^2	9	4	1	0	1	4	9

7. Find the coefficient of correlation between x and y for the following data:

x	62	64	65	69	70	71	72	74
y	126	125	139	145	165	152	180	208

[Ans.: 0.9032]

8. The following data gave the growth of employment in lacs in the organized sector in India between 1988 and 1995:

Year	1988	1989	1990	1991	1992	1993	1994	1995
Public sector	98	101	104	107	113	120	125	128
Private sector	65	65	67	68	68	69	68	68

Find the correlation coefficient between the employment in public and private sectors.

[Ans.: 0.77]

9. Calculate Karl Pearson's coefficient of correlation from the following data, using 20 as the working mean for price and 70 as working mean for demand.

Price	14	16	17	18	19	20	21	22	23
Demand	84	78	70	75	66	67	62	58	60

[Ans.: −0.954]

10. A sample of 25 pairs of values x and y lead to the following results:

$$\sum x=127, \sum y=100, \sum x^2=760, \sum y^2=449, \sum xy=500$$

Later on, it was found that two pairs of values were taken as (8, 14) and (8, 6) instead of the correct values (8, 12) and (6, 8). Find the corrected coefficient between x and y.

[Ans.: −0.31]

4.9 RANK CORRELATION

Let a group of n individuals be arranged in order of merit with respect to some characteristics. The same group would give a different order (rank) for different characteristics. Considering the orders corresponding to two characteristics A and B, the correlation between these n pairs of ranks is called the *rank correlation* in the characteristics A and B for that group of individuals.

4.9.1 Spearman's Rank Correlation Coefficient

Let x, y be the ranks of the i^{th} individuals in two characteristics A and B respectively where $i = 1, 2, \ldots, n$. Assuming that no two individuals have the same rank either for x or y, each of the variables x and y take the values $1, 2, \ldots, n$.

$$\bar{x}=\bar{y}=\frac{1+2+3+\cdots+n}{n}=\frac{n(n+1)}{2n}=\frac{n+1}{2}$$

$$\begin{aligned}
\sum(x-\bar{x})^2 &= \sum(x^2-2x\bar{x}+\bar{x}^2)\\
&= \sum x^2-2\bar{x}\sum x+\bar{x}^2\sum 1\\
&= \sum x^2-2n\bar{x}^2+n\bar{x}^2 \qquad \left[\because \sum x=n\bar{x} \text{ and } \sum 1=n\right]\\
&= \sum x^2-n\bar{x}^2\\
&= (1^2+2^2+\cdots+n^2)-n\left(\frac{n+1}{2}\right)^2\\
&= \frac{n(n+1)(2n+1)}{6}-\frac{n(n+1)^2}{4}\\
&= \frac{1}{12}(n^3-n)
\end{aligned}$$

Similarly, $\sum(y-\bar{y})^2=\frac{1}{12}(n^3-n)$

If d denotes the difference between the ranks of the i^{th} individuals in the two variables,

$$d=x-y=(x-\bar{x})-(y-\bar{y}) \qquad [\because \bar{x}=\bar{y}]$$

Squaring and summing over i from 1 to n,

$$\begin{aligned}
\sum d^2 &= \sum[(x-\bar{x})-(y-\bar{y})]^2\\
&= \sum(x-\bar{x})^2+\sum(y-\bar{y})^2-2\sum(x-\bar{x})(y-\bar{y})
\end{aligned}$$

$$\begin{aligned}
\sum(x-\bar{x})(y-\bar{y}) &= \frac{1}{2}\left[\sum(x-\bar{x})^2+\sum(y-\bar{y})^2-\sum d^2\right]\\
&= \frac{1}{12}(n^3-n)-\frac{1}{2}\sum d^2
\end{aligned}$$

Hence, the coefficient of correlation between these variables is

$$\begin{aligned}
r &= \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}\\
&= \frac{\frac{1}{12}(n^3-n)-\frac{1}{2}\sum d^2}{\frac{1}{12}(n^3-n)}\\
&= 1-\frac{6\sum d^2}{n^3-n}\\
&= 1-\frac{6\sum d^2}{n(n^2-1)}
\end{aligned}$$

This is called Spearman's rank correlation coefficient and is denoted by ρ.

Note $\sum d=\sum(x-y)=\sum x-\sum y=n(\bar{x}-\bar{y})=0$

Example 1

Ten participants in a contest are ranked by two judges as follows:

x	1	3	7	5	4	6	2	10	9	8
y	3	1	4	5	6	9	7	8	10	2

Calculate the rank correlation coefficient.

Solution

$n = 10$

Rank by first Judge x	Rank by second Judge y	$d = x - y$	d^2
1	3	−2	4
3	1	2	4
7	4	3	9
5	5	0	0
4	6	−2	4
6	9	−3	9
2	7	−5	25
10	8	2	4
9	10	−1	1
8	2	6	36
		$\sum d = 0$	$\sum d^2 = 96$

$$r = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

$$= 1 - \frac{6(96)}{10\left[(10)^2-1\right]}$$

$$= 0.418$$

Example 2

Ten competitors in a musical test were ranked by the three judges A, B, and C in the following order:

Rank by A	1	6	5	10	3	2	4	9	7	8
Rank by B	3	5	8	4	7	10	2	1	6	9
Rank by C	6	4	9	8	1	2	3	10	5	7

Using the rank correlation method, find which pair of judges has the nearest approach to common liking in music. **[Summer 2015]**

Solution

$n = 10$

Rank by A x	Rank by B y	Rank by C z	$d_1 = x - y$	$d_2 = y - z$	$d_3 = z - x$	d_1^2	d_2^2	d_3^2
1	3	6	−2	−3	5	4	9	25
6	5	4	1	1	−2	1	1	4
5	8	9	−3	−1	4	9	1	16
10	4	8	6	−4	−2	36	16	4
3	7	1	−4	6	−2	16	36	4
2	10	2	−8	8	0	64	64	0
4	2	3	2	−1	−1	4	1	1
9	1	10	8	−9	1	64	81	1
7	6	5	1	1	−2	1	1	4
8	9	7	−1	2	−1	1	4	1
			$\sum d_1 = 0$	$\sum d_2 = 0$	$\sum d_3 = 0$	$\sum d_1^2 = 200$	$\sum d_2^2 = 214$	$\sum d_3^2 = 60$

$$r(x, y) = 1 - \frac{6\sum d_1^2}{n(n^2-1)}$$

$$= 1 - \frac{6(200)}{10\left[(10)^2-1\right]}$$

$$= -0.21$$

$$r(y, z) = 1 - \frac{6\sum d_2^2}{n(n^2-1)}$$

$$= 1 - \frac{6(214)}{10\left[(10)^2-1\right]}$$

$$= -0.296$$

$$r(z, x) = 1 - \frac{6\sum d_3^2}{n(n^2-1)}$$

$$= 1 - \frac{6(60)}{10\left[(10)^2-1\right]}$$

$$= 0.64$$

Since $r(z, x)$ is maximum, the pair of judges A and C has the nearest common approach.

Example 3

Ten students got the following percentage of marks in mathematics and physics:

Mathematics (x)	8	36	98	25	75	82	92	62	65	35
Physics (y)	84	51	91	60	68	62	86	58	35	49

Find the rank correlation coefficient.

Solution

$n = 10$

x	y	Rank in mathematics x	Rank in Physics y	$d = x - y$	d^2
8	84	10	3	7	49
36	51	7	8	−1	1
98	91	1	1	0	0
25	60	9	6	3	9
75	68	4	4	0	0
82	62	3	5	−2	4
92	86	2	2	0	0
62	58	6	7	−1	1
65	35	5	10	−5	25
35	49	8	9	−1	1
				$\sum d = 0$	$\sum d^2 = 90$

$$r = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

$$= 1 - \frac{6(90)}{10\left[(10)^2 - 1\right]}$$

$$= 0.455$$

Example 4

The coefficient of rank correlation of the marks obtained by 10 students in physics and chemistry was found to be 0.5. It was later discovered that the difference in ranks in the two subjects obtained by one of the students was wrongly taken as 3 instead of 7. Find the rank coefficient of the rank correlation.

Solution

$n = 10$

$$r = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

$$0.5 = 1 - \frac{6\sum d^2}{10(100-1)}$$

$$\therefore \quad \sum d^2 = 82.5$$

$$\text{Correct} \sum d^2 = \text{Incorrect} \sum d^2 - (\text{Incorrect rank difference})^2 + (\text{Correct rank difference})^2$$

$$= 82.5 - (3)^2 + (7)^2$$

$$= 122.5$$

Correct coefficient of rank correlation $r = 1 - \frac{6(122.5)}{10(100-1)}$

$$= 0.26$$

4.9.2 Tied Ranks

If there is a tie between two or more individuals ranks, the rank is divided among equal individuals, e.g., if two items have fourth rank, the 4th and 5th rank is divided between them equally and is given as $\frac{4+5}{2} = 4.5^{\text{th}}$ rank to each of them. If three items have the same 4th rank, each of them is given $\frac{4+5+6}{3} = 5^{\text{th}}$ rank. As a result of this, the following adjustment or correction is made in the rank correlation formula. If m is the number of item having equal ranks then the factor $\frac{1}{12}(m^3 - m)$ is added to $\sum d^2$. If there are more than one cases of this type, this factor is added corresponding to each case.

$$r = 1 - \frac{6\left[\sum d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \cdots\right]}{n(n^2-1)}$$

Example 1

Obtain the rank correlation coefficient from the following data:

x	10	12	18	18	15	40
y	12	18	25	25	50	25

Solution

Here, $n = 6$

x	y	Rank x	Rank y	$d=x-y$	d^2
10	12	1	1	0	0
12	18	2	2	0	0
18	25	4.5	4	0.5	0.25
18	25	4.5	4	0.5	0.25
15	50	3	6	–3	9
40	25	6	4	2	4
					$\sum d^2 = 13.5$

There are two items in the x series having equal values at the rank 4. Each is given the rank 4.5. Similarly, there are three items in the y series at the rank 3. Each of them is given the rank 4.

$$m_1 = 2, m_2 = 3$$

$$r = 1 - \frac{6\left[\sum d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2)\right]}{n(n^2-1)}$$

$$= 1 - \frac{6\left[13.50 + \frac{1}{12}(8-2) + \frac{1}{12}(27-3)\right]}{6\left[(6)^2 - 1\right]}$$

$$= 0.5429$$

EXERCISE 4.2

1. Compute Spearman's rank correlation coefficient from the following data:

x	18	20	34	52	12
y	39	23	35	18	46

[Ans.: –0.9]

2. Two judges gave the following ranks to a series of eight one-act plays in a drama competition. Examine the relationship between their judgements.

Judge A	8	7	6	3	2	1	5	4
Judge B	7	5	4	1	3	2	6	8

[Ans.: 0.62]

3. From the following data, calculate Spearman's rank correlation between x and y.

x	36	56	20	42	33	44	50	15	60
y	50	35	70	58	75	60	45	80	38

[Ans.: 0.92]

4. Ten competitors in a voice test are ranked by three judges in the following order:

Rank by First Judge	6	10	2	9	8	1	5	3	4	7
Rank by Second Judge	5	4	10	1	9	3	8	7	2	6
Rank by Third Judge	4	8	2	10	7	6	9	1	3	6

Use the method of rank correlation to gauge which pairs of judges has the nearest approach to common liking in voice.

[Ans.: The first and third judge**]**

5. The following table gives the scores obtained by 11 students in English and Tamil translation. Find the rank correlation coefficient.

Scores in English	40	46	54	60	70	80	82	85	85	90	95
Scores in Tamil	45	45	50	43	40	75	55	72	65	42	70

[Ans.: 0.36]

6. Calculate Spearman's coefficient of rank correlation for the following data:

x	53	98	95	81	75	71	59	55
y	47	25	32	37	30	40	39	45

[Ans.: –0.905]

7. Following are the scores of ten students in a class and their IQ:

Score	35	40	25	55	85	90	65	55	45	50
IQ	100	100	110	140	150	130	100	120	140	110

Calculate the rank correlation coefficient between the score IQ.

[Ans.: 0.47]

4.10 REGRESSION

Regression is defined as a method of estimating the value of one variable when that of the other is known and the variables are correlated. *Regression analysis* is used to predict or estimate one variable in terms of the other variable. It is a highly valuable tool for prediction purpose in economics and business. It is useful in statistical estimation of demand curves, supply curves, production function, cost function, consumption function, etc.

4.11 TYPES OF REGRESSION

Regression is classified into two types:

1. Simple and multiple regressions
2. Linear and nonlinear regressions

4.11.1 Simple and Multiple Regressions

Depending upon the study of the number of variables, regression may be simple or multiple.

1. Simple Regression The regression analysis for studying only two variables at a time is known as simple regression.

2. Multiple Regression The regression analysis for studying more than two variables at a time is known as multiple regression.

4.11.2 Linear and Nonlinear Regressions

Depending upon the regression curve, regression may be linear or nonlinear.

1. Linear Regression If the regression curve is a straight line, the regression is said to be linear.

2. Nonlinear Regression If the regression curve is not a straight line i.e., not a first-degree equation in the variables x and y, the regression is said to be nonlinear or curvilinear. In this case, the regression equation will have a functional relation between the variables x and y involving terms in x and y of the degree higher than one, i.e., involving terms of the type x^2, y^2, x^3, y^3, xy, etc.

4.12 METHODS OF STUDYING REGRESSION

There are two methods of studying correlation:

(i) Method of scatter diagram
(ii) Method of least squares

4.12.1 Method of Scatter Diagram

It is the simplest method of obtaining the lines of regression. The data are plotted on a graph paper by taking the independent variable on the x-axis and the dependent variable on the y-axis. Each of these points are generally scattered in a narrow strip. If the correlation is perfect, i.e., if r is equal to one, positive, or negative, the points will lie on a line which is the line of regression.

4.12.2 Method of Least Squares

This is a mathematical method which gives an objective treatment to find a line of regression. It is used for obtaining the equation of a curve which fits best to a given set of observations. It is based on the assumption that the sum of squares of differences between the estimated values and the actual observed values of the observations is minimum.

4.13 LINES OF REGRESSION

If the variables, which are highly correlated, are plotted on a graph then the points lie in a narrow strip. If all the points in the scatter diagram cluster around a straight line, the line is called the *line of regression*. The line of regression is the line of best fit and is obtained by the principle of least squares.

Line of Regression of y on x

It is the line which gives the best estimate for the values of y for any given values of x. The regression equation of y on x is given by

$$y-\overline{y}=r\frac{\sigma_y}{\sigma_x}(x-\overline{x})$$

It is also written as

$$y=a+bx$$

Line of Regression of x on y

It is the line which gives the best estimate for the values of x for any given values of y. The regression equation for x on y is given by

$$x-\overline{x}=r\frac{\sigma_x}{\sigma_y}(y-\overline{y})$$

It is also written as

$$x=a+by$$

where $\overline{x}$ and $\overline{y}$ are means of x series and y series respectively, σ_x and σ_y are standard deviations of x series and y series respectively, r is the correlation coefficient between x and y.

4.14 REGRESSION COEFFICIENTS

The slope b of the line of regression of y on x is also called the *coefficient of regression* of y on x. It represents the increment in the value of y corresponding to a unit change in the value of x.

$$\begin{aligned} b_{yx} &= \text{Regression coefficient of } y \text{ on } x \\ &= r\frac{\sigma_y}{\sigma_x} \end{aligned}$$

Similarly, the slope b of the line of regression of x on y is called the coefficient of regression of x on y. It represents the increment in the value of x corresponding to a unit change in the value of y.

b_{xy} = Regression coefficient of x on y

$$= r\frac{\sigma_x}{\sigma_y}.$$

Expressions for Regression Coefficients

(i) We know that

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}$$

$$\sigma_x = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$$

$$\sigma_y = \sqrt{\frac{\sum(y-\bar{y})^2}{n}}$$

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$$

and $$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(y-\bar{y})^2}$$

(ii) We know that

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}\sqrt{\sum y^2 - \frac{\left(\sum y\right)^2}{n}}}$$

$$\sigma_x = \sqrt{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}$$

$$\sigma_y = \sqrt{\sum y^2 - \frac{\left(\sum y\right)^2}{n}}$$

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}$$

and $$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum y^2 - \frac{\left(\sum y\right)^2}{n}}$$

(iii) We know that

$$r = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sqrt{\sum d_x^2 - \frac{\left(\sum d_x\right)^2}{n}}\sqrt{\sum d_y^2 - \frac{\left(\sum d_y\right)^2}{n}}}$$

$$\sigma_x = \sqrt{\sum d_x^2 - \frac{\left(\sum d_x\right)^2}{n}}$$

$$\sigma_y = \sqrt{\sum d_y^2 - \frac{\left(\sum d_y\right)^2}{n}}$$

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sqrt{\sum d_x^2 - \frac{\left(\sum d_x\right)^2}{n}}}$$

and $$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sqrt{\sum d_y^2 - \frac{\left(\sum d_y\right)^2}{n}}}$$

4.15 PROPERTIES OF REGRESSION COEFFICIENTS

1. The coefficient of correlation is the geometric mean of the coefficients of regression, i.e., $r = \sqrt{b_{yx} b_{xy}}$.

Proof We know that

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

$$b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

$$b_{yx} b_{xy} = r\frac{\sigma_y}{\sigma_x} \cdot r\frac{\sigma_x}{\sigma_y}$$

$$= r^2$$

$$r = \sqrt{b_{yx} b_{xy}}$$

2. If one of the regression coefficients is greater than one, the other must be less than one.

Proof Let $b_{yx} > 1$

We know that

$$r^2 \le 1 \text{ and } r^2 = b_{yx} b_{xy}$$

$$b_{yx} b_{xy} \le 1$$

$$b_{yx} \le \frac{1}{b_{xy}}$$

Hence, if $b_{yx} < 1$, $b_{xy} > 1$

3. The arithmetic mean of regression coefficients is greater than or equal to the coefficient of correlation.

Proof We have to prove that

$$\frac{1}{2}(b_{yx} + b_{xy}) \ge r$$

i.e., $$\frac{1}{2}\left(r\frac{\sigma_y}{\sigma_x} + r\frac{\sigma_x}{\sigma_y}\right) \ge r$$

i.e., $$\frac{\sigma_y}{\sigma_x} + \frac{\sigma_x}{\sigma_y} \ge 2$$

i.e., $$\sigma_y^2 + \sigma_x^2 - 2\sigma_x\sigma_y \ge 0$$

i.e., $$(\sigma_y - \sigma_x)^2 \ge 0$$

which is always true, since the square of a real quantity is $1 \ge 0$.

4. Regression Coefficients are independent of the change of origin but not of scale.

Proof Let $$d_x = \frac{x-a}{h},\ d_y = \frac{y-b}{k}$$

$$x = a + hd_x,\ y = b + kd_y$$

where a, b, h (> 0) and k(> 0) are constants.

$$r_{d_x d_y} = r_{xy}, \sigma_{d_x}^2 = \frac{1}{h^2}\sigma_x^2, \sigma_{d_y}^2 = \frac{1}{k^2}\sigma_y^2$$

$$b_{d_x d_y} = r_{d_x d_y}\frac{\sigma_{d_x}}{\sigma_{d_y}}$$

$$= r_{xy}\frac{\sigma_x}{h}\frac{k}{\sigma_y}$$

$$= \frac{k}{h} r_{xy}\frac{\sigma_x}{\sigma_y}$$

$$= \frac{k}{h} b_{xy}$$

Similarly, $$b_{d_y d_x} = \frac{h}{k} b_{yx}$$

5. Both regression coefficients will have the same sign i.e., either both are positive or both are negative.

6. The sign of correlation is same as that of the regression coefficients, i.e., $r > 0$ if $b_{xy} > 0$ and $b_{yx} > 0$; and $r < 0$ if $b_{xy} < 0$ and $b_{yx} < 0$.

4.16 PROPERTIES OF LINES OF REGRESSION (LINEAR REGRESSION)

1. The two regression lines x on y and y on x always intersect at their means $(\overline{x}, \overline{y})$.
2. Since $r^2 = b_{yx} b_{xy}$, i.e., $r = \sqrt{b_{yx} b_{xy}}$, therefore, r, b_{yx}, b_{xy} all have the same sign.
3. If $r = 0$, the regression coefficients are zero.
4. The regression lines become identical if $r = \pm 1$. It follows from the regression equations that $x = \overline{x}$ and $y = \overline{y}$. If $r = 0$, these lines are perpendicular to each other.

Example 1

The regression lines of a sample are $x + 6y = 6$ and $3x + 2y = 10$. Find
(i) sample means $\bar{x}$ and $\bar{y}$, and
(ii) the coefficient of correlation between x and y.
(iii) Also estimate y when $x = 12$.

Solution

(i) The regression lines pass through the point $(\bar{x}, \bar{y})$.

$$\bar{x} + 6\bar{y} = 6 \qquad \text{...(1)}$$

$$3\bar{x} + 2\bar{y} = 10 \qquad \text{...(2)}$$

Solving Eqs (1) and (2),

$$\bar{x} = 3, \quad \bar{y} = \frac{1}{2}$$

(ii) Let the line $x + 6y = 6$ be the line of regression of y on x.

$$6y = -x + 6$$

$$y = -\frac{1}{6}x + 1$$

$$\therefore \quad b_{yx} = -\frac{1}{6}$$

Let the line $3x + 2y = 10$ be the line of regression of x on y.

$$3x = -2y + 10$$

$$x = -\frac{2}{3}y + \frac{10}{3}$$

$$\therefore \quad b_{xy} = -\frac{2}{3}$$

$$r = \sqrt{b_{yx}\, b_{xy}} = \sqrt{\left(-\frac{1}{6}\right)\left(-\frac{2}{3}\right)} = \frac{1}{3}$$

Since b_{yx} and b_{xy} are negative, r is negative.

$$r = -\frac{1}{3}$$

Estimated value of y when $x = 12$ is

$$y = -\frac{1}{6}(12) + 1 = -1$$

Example 2

If the two lines of regression are $4x - 5y + 30 = 0$ and $20x - 9y - 107 = 0$, which of these are lines of regression of x on y and y on x? Find r_{xy} and σ_y when $\sigma_x = 3$.

Solution

For the line $\quad 4x - 5y + 30 = 0,$

$$-5y = -4x - 30$$

$$y = 0.8x + 6$$

$$\therefore \quad b_{yx} = 0.8$$

For the line $\quad 20x - 9y - 107 = 0$

$$20x = 9y + 107$$

$$x = 0.45y + 5.35$$

$$\therefore \quad b_{xy} = 0.45$$

Both b_{yx} and b_{xy} are positive.
Hence, line $4x - 5y + 30 = 0$ is the line of regression of y one x and line $20x - 9y - 107 = 0$ is the line of regression of x on y.

$$r = \sqrt{b_{yx}\, b_{xy}} = \sqrt{(0.8)(0.45)} = 0.6$$

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

$$0.8 = 0.6\left(\frac{\sigma_y}{3}\right)$$

$$\therefore \quad \sigma_y = 4$$

Example 3

The following data regarding the heights (y) and weights (x) of 100 college students are given:

$$\sum x = 15000, \quad \sum x^2 = 2272500, \quad \sum y = 6800$$

$$\sum y^2 = 463025, \quad \sum xy = 1022250$$

Find the coefficient of correlation between height and weight and also the equation of regression of height and weight.

Solution

$$n = 100$$

$$b_{yx} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}$$

$$= \frac{1022250 - \frac{(15000)(6800)}{100}}{2272500 - \frac{(15000)^2}{100}}$$

$$= 0.1$$

$$b_{xy} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum y^2 - \frac{\left(\sum y\right)^2}{n}}$$

$$= \frac{1022250 - \frac{(15000)(6800)}{100}}{463025 - \frac{(6800)^2}{100}}$$

$$= 3.6$$

$$r = \sqrt{b_{yx}\, b_{xy}} = \sqrt{(0.1)(3.6)} = 0.6$$

$$\bar{x} = \frac{\sum x}{n} = \frac{15000}{100} = 150$$

$$\bar{y} = \frac{\sum y}{n} = \frac{6800}{100} = 68$$

The equation of the line of regression of y on x is

$$y - \bar{y} = b_{yx}\,(x - \bar{x})$$

$$y - 68 = 0.1(x - 150)$$

$$y = 0.1x + 53$$

The equation of the line of regression of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 150 = 3.6(y - 68)$$

$$x = 3.6y - 94.8$$

Example 4

For a bivariate data, the mean value of x is 20 and the mean value of y is 45. The regression coefficient of y on x is 4 and that of x on y is $\frac{1}{9}$. Find

(i) the coefficient of correlation, and

(ii) the standard deviation of x if the standard deviation of y is 12.

(iii) Also write down the equations of regression lines..

Solution

$$\bar{x} = 20, \quad \bar{y} = 45, \quad b_{yx} = 4, \quad b_{xy} = \frac{1}{9}$$

(i) $r = \sqrt{b_{yx}\, b_{xy}} = \sqrt{(4)\left(\frac{1}{9}\right)} = \frac{2}{3} = 0.667$

(ii)

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

$$4 = \frac{2}{3}\left(\frac{12}{\sigma_x}\right)$$

$$\therefore \quad \sigma_x = 2$$

(iii) The equation of the regression line of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 45 = 4(x - 20)$$

$$y = 4x - 35$$

The equation of the regression line of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 20 = \frac{1}{9}(y - 45)$$

$$x = \frac{1}{9}y + 15$$

Example 5

From the following results, obtain the two regression equations and estimate the yield when the rainfall is 29 cm and the rainfall, when the yield is 600 kg:

	Yield in kg	Rainfall in cm
Mean	508.4	26.7
SD	36.8	4.6

The coefficient of correlation between yield and rainfall is 0.52.

Solution

Let rainfall in cm be denoted by x and yield in kg be denoted by y.

$$\bar{x} = 26.7, \quad \bar{y} = 508.4, \quad \sigma_x = 4.6, \quad \sigma_y = 36.8, \quad r = 0.52$$

$$\begin{aligned} b_{yx} &= r\frac{\sigma_y}{\sigma_x} \\ &= 0.52\left(\frac{36.8}{4.6}\right) \\ &= 4.16 \end{aligned}$$

$$\begin{aligned} b_{xy} &= r\frac{\sigma_x}{\sigma_y} \\ &= 0.52\left(\frac{4.6}{36.8}\right) \\ &= 0.065 \end{aligned}$$

The equation of the line of regression of y on x is

$$\begin{aligned} y - \bar{y} &= b_{yx}(x - \bar{x}) \\ y - 508.4 &= 4.16(x - 26.7) \\ y &= 4.16x + 397.328 \end{aligned}$$

The equation of the line of regression of x on y is

$$\begin{aligned} x - \bar{x} &= b_{xy}(y - \bar{y}) \\ x - 26.7 &= 0.065(y - 508.4) \\ x &= 0.065y - 6.346 \end{aligned}$$

Estimated yield when the rainfall is 29 cm is

$$y = 4.16\,(29) + 397.328 = 517.968 \text{ kg}$$

Estimated rainfall when the yield is 600 kg is

$$x = 0.065\,(600) - 6.346 = 32.654 \text{ cm}$$

Example 6

Find the regression coefficients b_{yx} *and* b_{xy} *and hence, find the correlation coefficient between x and y for the following data:*

x	4	2	3	4	2
y	2	3	2	4	4

Solution

$n = 5$

x	y	x^2	y^2	xy
4	2	16	4	8
2	3	4	9	6
3	2	9	4	6
4	4	16	16	16
2	4	4	16	8
$\Sigma x = 15$	$\Sigma y = 15$	$\Sigma x^2 = 49$	$\Sigma y^2 = 49$	$\Sigma xy = 44$

$$\begin{aligned} b_{yx} &= \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}} \\ &= \frac{44 - \frac{(15)(15)}{5}}{49 - \frac{(15)^2}{5}} \\ &= -0.25 \end{aligned}$$

$$\begin{aligned} b_{xy} &= \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum y^2 - \frac{\left(\sum y\right)^2}{n}} \\ &= \frac{44 - \frac{(15)(15)}{5}}{49 - \frac{(15)^2}{5}} \\ &= -0.25 \end{aligned}$$

$$r = \sqrt{b_{yx}\, b_{xy}} = \sqrt{(-0.25)(-0.25)} = 0.25$$

Since b_{yx} and b_{xy} are negative, r is negative.

$$r = -0.25$$

Note $\sum x, \sum y, \sum x^2, \sum y^2, \sum xy$ can be directly obtained with the help of scientific calculator.

Example 7

The following data give the experience of machine operators and their performance rating as given by the number of good parts turned out per 100 pieces.

Operator	1	2	3	4	5	6
Performance rating (x)	23	43	53	63	73	83
Experience (y)	5	6	7	8	9	10

Calculate the regression line of performance rating on experience and also estimate the probable performance if an operator has 11 years of experience. **[Summer 2015]**

Solution

$n = 6$

x	y	y^2	xy
23	5	25	115
43	6	36	258
53	7	49	371
63	8	64	504
73	9	81	657
83	10	100	830
$\sum x = 338$	$\sum y = 45$	$\sum y^2 = 355$	$\sum xy = 2735$

$$b_{xy} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum y^2 - \frac{\left(\sum y\right)^2}{n}}$$

$$= \frac{2735 - \frac{(338)(45)}{6}}{355 - \frac{(45)^2}{6}}$$

$$= 11.429$$

$$\bar{x} = \frac{\sum x}{n} = \frac{338}{6} = 56.33$$

$$\bar{y} = \frac{\sum y}{n} = \frac{45}{6} = 7.5$$

The equation of regression line of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 56.33 = 11.429(y - 7.5)$$

$$x = 11.429\,y - 29.3875$$

Estimated performance if $y = 11$ is

$$x = 11.429(11) - 29.3875 = 96.3315$$

Example 8

The number of bacterial cells (y) per unit volume in a culture at different hours (x) is given below:

x	0	1	2	3	4	5	6	7	8	9
y	43	46	82	98	123	167	199	213	245	272

Fit lines of regression of y on x and x on y. Also, estimate the number of bacterial cells after 15 hours.

Solution

$n = 10$

x	y	x^2	xy	y^2
0	43	0	0	1849
1	46	1	46	2116
2	82	4	164	6724
3	98	9	294	9604
4	123	16	492	15129
5	167	25	835	27889
6	199	36	1194	39601
7	213	49	1491	45369
8	245	64	1960	60025
9	272	81	2448	73984
$\sum x = 45$	$\sum y = 1488$	$\sum x^2 = 285$	$\sum xy = 8924$	$\sum y^2 = 282290$

$$b_{yx} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}$$

$$= \frac{8924 - \frac{(45)(1488)}{10}}{285 - \frac{(45)^2}{10}}$$

$$= 27.0061$$

$$b_{xy} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum y^2 - \frac{\left(\sum y\right)^2}{n}}$$

$$= \frac{8924 - \frac{(45)(1488)}{10}}{282290 - \frac{(1488)^2}{10}}$$

$$= 0.0366$$

$$\overline{x} = \frac{\sum x}{n} = \frac{45}{10} = 4.5$$

$$\overline{y} = \frac{\sum y}{n} = \frac{1488}{10} = 148.8$$

The equation of the line of regression of y on x is

$$y - \overline{y} = b_{yx}(x - \overline{x})$$

$$y - 148.8 = 27.0061\,(x - 4.5)$$

$$y = 27.0061x + 27.2726$$

The equation of the line of regression of x on y is

$$x - \overline{x} = b_{xy}(y - \overline{y})$$

$$x - 4.5 = 0.0366(y - 148.8)$$

$$x = 0.366\,y - 0.9461$$

At $x = 15$ hours,

$$y = 27.0061\ (15) + 27.2726 = 432.3641$$

Example 9

Find the regression coefficient of y on x for the following data:

x	1	2	3	4	5
y	160	180	140	180	200

Solution

$$n = 5$$

$$\overline{x} = \frac{\sum x}{n} = \frac{15}{5} = 3$$

$$\overline{y} = \frac{\sum y}{n} = \frac{860}{5} = 172$$

x	y	$x-\overline{x}$	$y-\overline{y}$	$(x-\overline{x})^2$	$(x-\overline{x})(y-\overline{y})$
1	160	–2	–12	4	24
2	180	–1	8	1	–8
3	140	0	–32	0	0
4	180	1	8	1	8
5	200	2	28	4	56
$\sum x = 15$	$\sum y = 860$	$\sum(x-\overline{x}) = 0$	$\sum(y-\overline{y}) = 0$	$\sum(x-\overline{x})^2 = 10$	$\sum(x-\overline{x})(y-\overline{y}) = 80$

$$b_{yx} = \frac{\sum (x-\overline{x})(y-\overline{y})}{\sum (x-\overline{x})^2}$$

$$= \frac{80}{10}$$

$$= 8$$

Note Since $\sum x, \sum y, \sum x^2, \sum y^2, \sum xy$ can be directly obtained with the help of scientific calculator, the regression coefficient can be calculated without using mean.

Example 10

Calculate the two regression coefficients from the data and find correlation coefficient.

x	7	4	8	6	5
y	6	5	9	8	2

Solution

$$n = 5$$

$$\overline{x} = \frac{\sum x}{n} = \frac{30}{5} = 6$$

$$\overline{y} = \frac{\sum y}{n} = \frac{30}{5} = 6$$

x	y	$x-\overline{x}$	$y-\overline{y}$	$(x-\overline{x})^2$	$(y-\overline{y})^2$	$(x-\overline{x})(y-\overline{y})$
7	6	1	0	1	0	0
4	5	−2	−1	4	1	2
8	9	2	3	4	9	6
6	8	0	2	0	4	0
5	2	−1	−4	1	16	4
$\sum x = 30$	$\sum y = 30$	$\sum(x-\overline{x}) = 0$	$\sum(y-\overline{y}) = 0$	$\sum(x-\overline{x})^2 = 10$	$\sum(y-\overline{y})^2 = 30$	$\sum(x-\overline{x})(y-\overline{y}) = 12$

$$\begin{aligned} b_{yx} &= \frac{\sum(x-\overline{x})(y-\overline{y})}{\sum(x-\overline{x})^2} \\ &= \frac{12}{10} \\ &= 1.2 \end{aligned}$$

$$\begin{aligned} b_{xy} &= \frac{\sum(x-\overline{x})(y-\overline{y})}{\sum(y-\overline{y})^2} \\ &= \frac{12}{30} \\ &= 0.4 \end{aligned}$$

$$r = \sqrt{b_{yx}\, b_{xy}} = \sqrt{(1.2)(0.4)} = 0.693$$

Example 11

Obtain the two regression lines from the following data and hence, find the correlation coefficient.

x	6	2	10	4	8
y	9	11	5	8	7

[Summer 2015]

Solution

$$n = 5$$

$$\overline{x} = \frac{\sum x}{n} = \frac{30}{5} = 6$$

$$\overline{y} = \frac{\sum y}{n} = \frac{40}{5} = 8$$

x	y	$x-\overline{x}$	$y-\overline{y}$	$(x-\overline{x})^2$	$(y-\overline{y})^2$	$(x-\overline{x})(y-\overline{y})$
6	9	0	1	0	1	0
2	11	−4	3	16	9	−12
10	5	4	−3	16	9	−12
4	8	−2	0	4	0	0
8	7	2	−1	4	1	−2
$\sum x = 30$	$\sum y = 40$	$\sum(x-\overline{x}) = 0$	$\sum(y-\overline{y}) = 0$	$\sum(x-\overline{x})^2 = 40$	$\sum(y-\overline{y})^2 = 20$	$\sum(x-\overline{x})(y-\overline{y})^2 = -26$

$$\begin{aligned} b_{yx} &= \frac{\sum(x-\overline{x})(y-\overline{y})}{\sum(x-\overline{x})^2} \\ &= \frac{-26}{40} \\ &= -0.65 \end{aligned}$$

$$\begin{aligned} b_{xy} &= \frac{\sum(x-\overline{x})(y-\overline{y})}{\sum(y-\overline{y})^2} \\ &= \frac{-26}{20} \\ &= -1.3 \end{aligned}$$

The equation of regression line of y on x is

$$\begin{aligned} y-\overline{y} &= b_{yx}(x-\overline{x}) \\ y-8 &= -0.65(x-6) \\ y &= -0.65\,x + 11.9 \end{aligned}$$

The equation of regression line of x on y is

$$\begin{aligned} x-\overline{x} &= b_{xy}(y-\overline{y}) \\ x-6 &= -1.3(y-8) \\ x &= -1.3\,y + 16.4 \end{aligned}$$

$$r = \sqrt{b_{yx}\, b_{xy}} = \sqrt{(-0.65)(-1.3)} = 0.9192$$

Since b_{yx} and b_{xy} are negative, r is negative.

$$r = -0.9192.$$

Example 12

Calculate the regression coefficients and find the two lines of regression from the following data:

x	57	58	59	59	60	61	62	64
y	67	68	65	68	72	72	69	71

Find the value of y when x = 66.

Solution

$$n = 8$$

$$\bar{x} = \frac{\sum x}{n} = \frac{480}{8} = 60$$

$$\bar{y} = \frac{\sum y}{n} = \frac{552}{8} = 69$$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
57	67	−3	−2	9	4	6
58	68	−2	−1	4	1	2
59	65	−1	−4	1	16	4
59	68	−1	−1	1	1	1
60	72	0	3	0	9	0
61	72	1	3	1	9	3
62	69	2	0	4	0	0
64	71	4	2	16	4	8
$\Sigma x = 480$	$\Sigma y = 552$	$\Sigma(x-\bar{x}) = 0$	$\Sigma(y-\bar{y}) = 0$	$\Sigma(x-\bar{x})^2 = 36$	$\Sigma(y-\bar{y})^2 = 44$	$\Sigma(x-\bar{x})(y-\bar{y}) = 24$

$$b_{yx} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$$

$$= \frac{24}{36}$$

$$= 0.667$$

$$b_{xy} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (y-\bar{y})^2}$$

$$= \frac{24}{44}$$

$$= 0.545$$

The equation of regression line of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 69 = 0.667(x - 60)$$

$$y = 0.667x + 28.98$$

The equation of regression line of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 60 = 0.545(y - 69)$$

$$x = 0.545y + 22.395$$

Value of y when $x = 66$ is

$$y = 0.667\,(66) + 28.98 = 73.002$$

Example 13

The following data represents rainfall (x) and yield of paddy per hectare (y) in a particular area. Find the linear regression of x on y.

x	113	102	95	120	140	130	125
y	1.8	1.5	1.3	1.9	1.1	2.0	1.7

Solution

Let $a = 120$ and $b = 1.8$ be the assumed means of x and y series respectively.

$$d_x = x - a = x - 120$$

$$d_y = y - b = y - 1.8$$

$$n = 7$$

x	y	d_x	d_y	d_y^2	d_xd_y
113	1.8	–7	0	0	0
102	1.5	–18	–0.3	0.09	5.4
95	1.3	–25	–0.5	0.25	12.5
120	1.9	0	0.1	0.01	0
140	1.1	20	–0.7	0.49	–14
130	2.0	10	0.2	0.04	2.0
125	1.7	5	–0.1	0.01	–0.5
$\Sigma x = 825$	$\Sigma y = 11.3$	$\Sigma d_x = -15$	$\Sigma d_y = -1.3$	$\Sigma d_y^2 = 0.89$	$\Sigma d_xd_y = 5.4$

$$b_{xy} = \frac{\sum d_xd_y - \frac{\sum d_x \sum d_y}{n}}{\sum d_y^2 - \frac{\left(\sum d_y\right)^2}{n}}$$

$$= \frac{5.4 - \frac{(-15)(-1.3)}{7}}{0.89 - \frac{(-1.3)^2}{7}}$$

$$= 4.03$$

$$\overline{x} = \frac{\sum x}{n} = \frac{825}{7} = 117.86$$

$$\overline{y} = \frac{\sum y}{n} = \frac{11.3}{7} = 1.614$$

The equation of the regression line of x on y is

$$x - \overline{x} = b_{xy}(y - \overline{y})$$

$$x - 117.86 = 4.03\,(y - 1.614)$$

$$x = 4.03\,y + 111.36$$

Note Since Σx, Σy, Σx^2, Σy^2, Σxy can be directly obtained with the help of scientific calculator, the regression coefficient can be calculated without using assumed mean.

Example 14

Find the two lines of regression from the following data:

Age of husband (x)	25	22	28	26	35	20	22	40	20	18
Age of wife (y)	18	15	20	17	22	14	16	21	15	14

Hence, estimate (i) the age of the husband when the age of the wife is 19, and (ii) the age of the wife when the age of the husband is 30.

Solution

Let $a = 26$ and $b = 17$ be the assumed means of x and y series respectively.

$$d_x = x - a = x - 26$$

$$d_y = y - b = y - 17$$

$$n = 10$$

x	y	d_x	d_y	d_x^2	d_y^2	d_xd_y
25	18	–1	1	1	1	–1
22	15	–4	–2	16	4	8
28	20	2	3	4	9	6
26	17	0	0	0	0	0
35	22	9	5	81	25	45
20	14	–6	–3	36	9	18
22	16	–4	–1	16	1	4
40	21	14	4	196	16	56
20	15	–6	–2	36	4	12
18	14	–8	–3	64	9	24
$\Sigma x = 256$	$\Sigma y = 172$	$\Sigma d_x = -4$	$\Sigma d_y = 2$	$\Sigma d_x^2 = 450$	$\Sigma d_y^2 = 78$	$\Sigma d_xd_y = 172$

$$b_{yx} = \frac{\sum d_xd_y - \frac{\sum d_x \sum d_y}{n}}{\sum d_x^2 - \frac{\left(\sum d_x\right)^2}{n}}$$

$$= \frac{172 - \frac{(-4)(2)}{10}}{450 - \frac{(-4)^2}{10}}$$

$$= 0.385$$

$$b_{xy} = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sum d_y^2 - \frac{\left(\sum d_y\right)^2}{n}}$$

$$= \frac{172 - \frac{(-4)(2)}{10}}{78 - \frac{(2)^2}{10}}$$

$$= 2.227$$

$$\overline{x} = \frac{\sum x}{n} = \frac{256}{10} = 25.6$$

$$\overline{y} = \frac{\sum y}{n} = \frac{172}{10} = 17.2$$

The equation of the regression line of y on x is

$$y - \overline{y} = b_{yx}(x - \overline{x})$$

$$y - 17.2 = 0.385(x - 25.6)$$

$$y = 0.385x + 7.344$$

The equation of the regression line of x on y is

$$x - \overline{x} = b_{xy}(y - \overline{y})$$

$$x - 25.6 = 2.227(y - 17.2)$$

$$x = 2.227y - 12.704$$

Estimated age of the husband when the age of the wife is 19 is

$$x = 2.227\,(19) - 12.704 = 29.601 \text{ or } 30 \text{ nearly}$$

Age of the husband = 30 years

Estimated age of the wife when the age of the husband is 30 is

$$y = 0.385\,(30) + 7.344 = 18.894 \text{ or } 19 \text{ nearly}$$

Age of the wife = 19 years

Example 15

From the following data, obtain the two regression lines and correlation coefficient.

Sales (x)	100	98	78	85	110	93	80
Purchase (y)	85	90	70	72	95	81	74

Solution

Let $a = 93$ and $b = 81$ be the assumed means of x and y series respectively.

$$d_x = x - a = x - 93$$

$$d_y = y - b = y - 91$$

$$n = 7$$

x	y	d_x	d_y	d_x^2	d_y^2	$d_x d_y$
100	85	7	4	49	16	28
98	90	5	9	25	81	45
78	70	–15	–11	225	121	165
85	72	–8	–9	64	81	72
110	95	17	14	289	196	238
93	81	0	0	0	0	0
80	74	–13	–7	169	49	91
$\sum x = 644$	$\sum y = 567$	$\sum d_x = -7$	$\sum d_y = 0$	$\sum d_x^2 = 821$	$\sum d_y^2 = 544$	$\sum d_x d_y = 639$

$$b_{yx} = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sum d_x^2 - \frac{\left(\sum d_x\right)^2}{n}}$$

$$= \frac{639 - \frac{(-7)(0)}{7}}{821 - \frac{(-7)^2}{7}}$$

$$= 0.785$$

$$b_{xy} = \frac{\sum d_x dy - \frac{\sum d_x \sum d_y}{n}}{\sum d_y^2 - \frac{\left(\sum d_y\right)^2}{n}}$$

$$= \frac{639 - \frac{(-7)(0)}{7}}{544 - \frac{(0)^2}{7}}$$

$$= 1.1746$$

$$\bar{x} = \frac{\sum x}{n} = \frac{644}{7} = 92$$

$$\bar{y} = \frac{\sum y}{n} = \frac{567}{7} = 81$$

The equation of regression line of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 81 = 0.785(x - 92)$$

$$y = 0.785x + 8.78$$

The equation of regression line of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 92 = 1.1746(y - 81)$$

$$x = 1.1746 - 3.1426$$

$$r = \sqrt{b_{yx}\, b_{xy}} = \sqrt{(0.785)(1.1746)} = 0.9602$$

EXERCISE 4.3

1. The following are the lines of regression $4y = x + 38$ and $9y = x + 288$. Estimate y when $x = 99$ and x when $y = 30$. Also, find the means of x and y.

[**Ans.:** $y = 43, x = 82, \bar{x} = 162, \bar{y} = 50$]

2. The equations of the two lines of regression are $x = 19.13 - 0.87\, y$ and $y = 11.64 - 0.50\, x$. Find (i) the means of x and y, and (ii) the coefficient of correlation between x and y.

[**Ans.:** $\bar{x} = 15.79, \bar{y} = 3.74$, (ii) $r = -0.66, b_{yx} = -0.5, b_{xy} = 0.87$]

3. Given var(x) = 25. The equations of the two lines of regression are $5x - y = 22$ and $64\, x - 45\, y = 24$. Find (i) $\bar{x}$ and $\bar{y}$, (ii) r, and (iii) σ_y.

[**Ans.:** $\bar{x} = 6, \bar{y} = 8$, (ii) $r = 1.87$ (iii) $\sigma_y = 0.2$]

4. In a partially destroyed laboratory record of analysis of correlation data the following results are legible. Variance = 9, the equations of the lines of regression $4x - 5y + 33 = 0$, $20\, x - 9\, y - 107 = 0$. Find (i) the mean values of x and y, (ii) the standard deviation of y, and (iii) the coefficient of correlation between x and y

[**Ans.:** (i) $\bar{x} = 13, \bar{y} = 17$, (ii) $\sigma_y = 4$, (iii) $r = 0.6$]

5. From a sample of 200 pairs of observation, the following quantities were calculated:

$$\sum x = 11.34, \sum y = 20.78, \sum x^2 = 12.16, \sum y^2 = 84.96, \sum xy = 22.13$$

From the above data, show how to compute the coefficients of the equation $y = a + bx$.

[**Ans.:** $a = 0.0005, b = 1.82$]

6. In the estimation of regression equations of two variables x and y, the following results were obtained:

$$\bar{x} = 90, \bar{y} = 70, n = 10, \Sigma(x - \bar{x})^2 = 6360, \Sigma(y - \bar{y})^2 = 2860$$

$$\Sigma(x - \bar{x})(y - \bar{y}) = 3900$$

Obtain the two lines of regression.

[**Ans.:** $x = 1.361\, y - 5.27$, $y = 0.613\, x + 14.812$]

7. Find the likely production corresponding to a rainfall of 40 cm from the following data:

	Rainfall (in cm)	Output (in quintals)
mean	30	50
SD	5	10
$r = 0.8$		

[**Ans.:** 66 quintals]

8. The following table gives the age of a car of a certain make and annual maintenance cost. Obtain the equation of the line of regression of cost on age.

Age of a car	2	4	6	8
Maintenance	1	2	2.5	3

[**Ans.:** $x = 0.325\, y + 0.5$]

9. Obtain the equation of the line of regression of y on x from the following data and estimate y for $x = 73$.

x	70	72	74	76	78	80
y	163	170	179	188	196	220

[**Ans.:** $y = 5.31\, x - 212.57$, $y = 175.37$]

10. The heights in cm of fathers (x) and of the eldest sons (y) are given below:

x	165	160	170	163	173	158	178	168	173	170	175	180
y	173	168	173	165	175	168	173	165	180	170	173	178

Estimate the height of the eldest son if the height of the father is 172 cm and the height of the father if the height of the eldest son is 173 cm. Also, find the coefficient of correlation between the heights of fathers and sons.

[Ans.: (i) $y = 1.016\,x - 5.123$ (ii) $x = 0.476\,y + 98.98$
(iii) 169.97, 173.45 (iv) $r = 0.696$]

11. Find (i) the lines of regression, and (ii) coefficient of correlation for the following data:

x	65	66	67	67	68	69	70	72
y	67	68	65	66	72	72	69	71

[Ans.: (i) $y = 19.64 + 0.72\,x$, $x = 33.29 + 0.5\,y$, (ii) $r = 0.604$]

12. Find the line of regression for the following data and estimate y corresponding to $x = 15.5$.

x	10	12	13	16	17	20	25
y	19	22	24	27	29	33	37

[Ans.: $y = 1.21x + 7.71$, $y = 26.465$]

13. The following data give the heights in inches (x) and weights in lbs (y) of a random sample of 10 students:

x	61	68	68	64	65	70	63	62	64	67
y	112	123	130	115	110	125	100	113	116	126

Estimate the weight of a student of height 59 inches.

[Ans.: 126.4 lbs]

14. Find the regression equations of y on x from the data given below taking deviations from actual mean of x and y.

Price in rupees (x)	10	12	13	12	16	15
Demand (y)	40	38	43	45	37	43

Estimate the demand when the price is ₹20.

[Ans.: $y = -0.25\,x + 44.25$, $y = 39.25$]

CHAPTER 5

Some Special Probability Distributions

Chapter Outline

5.1 Introduction
5.2 Binomial Distribution
5.3 Poisson Distribution
5.4 Normal Distribution
5.5 Exponential Distribution
5.6 Gamma Distribution

5.1 INTRODUCTION

There are some specific distributions that are used in practice. There is a random experiment behind each of these distributions. Since these random experiments model a lot of real life phenomenon, these special distributions are used frequently in different applications. Often a random experiment that we encounter in practice is such that we are interested in the associated random variable X with such a standard distribution. This chapter discusses special random variables and their distributions. These include binomial distribution, Poisson distribution, normal distribution, exponential distribution and gamma distribution.

5.2 BINOMIAL DISTRIBUTION

Consider n independent trials of a random experiments which results in either success or failure. Let p be the probability of success remaining constant every time and $q = 1 - p$ be the probability of failure. The probability of x successes and $n - x$ failures is given by $p^x q^{n-x}$ (multiplication theorem of probability). But these x successes and $n - x$ failures can occur in any of the ${}^nC_x$ ways in each of which the probability is same. Hence, the probability of x successes is ${}^nC_x p^x q^{n-x}$.

$$P(X = x) = {}^nC_x\, p^x\, q^{n-x}, \quad x = 0, 1, 2, \ldots, n, \text{ where } p + q = 1$$

A random variable X is said to follow the binomial distribution if the probability of x is given by

$$P(X = x) = p(x) = {}^nC_x\, p^x\, q^{n-x}, \quad x = 0, 1, 2, \ldots, n \text{ and } q = 1 - p$$

The two constants n and p are called the parameters of the distribution.

5.2.1 Examples of Binomial Distribution

(i) Number of defective bolts in a box containing n bolts.
(ii) Number of post-graduates in a group of n people.
(iii) Number of oil wells yielding natural gas in a group of n wells test drilled.
(iv) Number of machines lying idle in a factory having n machines.

5.2.2 Conditions for Binomial Distribution

The binomial distribution holds under the following conditions:

(i) The number of trials n is finite.
(ii) There are only two possible outcomes, success or failure.
(iii) The trials are independent of each other.
(iv) The probability of success p is constant for each trial.

5.2.3 Constants of the Binomial Distribution

1. Mean of the Binomial Distribution

$$\begin{aligned}
E(X) &= \sum x\, p(x) \\
&= \sum_{x=0}^{n} x\, {}^nC_x\, p^x\, q^{n-x} \\
&= 0 \cdot {}^nC_0\, p^0\, q^n + 1 \cdot {}^nC_1\, p q^{n-1} + 2 \cdot {}^nC_2\, p^2\, q^{n-2} + \cdots + n\, p^n \\
&= np\,[q^{n-1} + {}^{(n-1)}C_1\, q^{n-2}\, p + {}^{(n-1)}C_2\, q^{n-3}\, p^2 + \cdots + p^{n-1}] \\
&= np\,(q + p)^{n-1} \\
&= np \qquad [\because\ p + q = 1]
\end{aligned}$$

2. Variance of the Binomial Distribution

$$\begin{aligned}
\text{Var}(X) &= E(X^2) - \mu^2 \\
&= \sum_{x=0}^{n} x^2\, p(x) - \mu^2 \\
&= \sum_{x=0}^{n} x^2\, {}^nC_x\, p^x\, q^{n-x} - \mu^2 \\
&= \sum_{x=0}^{n} [x + x(x-1)]\, {}^nC_x\, p^x\, q^{n-x} - \mu^2 \\
&= \sum_{x=0}^{n} x\, {}^nC_x\, p^x\, q^{n-x} + \sum_{x=0}^{n} x(x-1)\, {}^nC_x\, p^x\, q^{n-x} - \mu^2 \\
&= np + \sum_{x=0}^{n} x(x-1) \frac{n(n-1)}{x(x-1)} \cdot {}^{(n-2)}C_{x-2} p^x\, q^{n-x} - \mu^2 \\
&= np + \sum_{x=0}^{n} n(n-1) \cdot {}^{(n-2)}C_{x-2} p^2 p^{x-2}\, q^{n-x} - \mu^2 \\
&= np + n(n-1)\, p^2 \sum_{x=0}^{n} {}^{(n-2)}C_{x-2}\, p^{x-2}\, q^{n-x} - \mu^2 \\
&= np + n(n-1)p^2 \cdot (q + p)^{n-2} - \mu^2 \\
&= np + n(n-1)\, p^2 - \mu^2 \qquad [\because\ p + q = 1] \\
&= np\,[1 + (n-1)p] - \mu^2 \\
&= np\,[1 - p + np] - \mu^2 \\
&= np\,[q + np] - \mu^2 \qquad [\because\ 1 - p = q] \\
&= np\,(q + np) - (np)^2 \\
&= npq
\end{aligned}$$

3. Standard Deviation of the Binomial Distribution

$$\text{SD} = \sqrt{\text{Variance}} = \sqrt{npq}$$

4. Mode of the Binomial Distribution

Mode of the binomial distribution is the value of x at which $p(x)$ has maximum value.

Mode = integral part of $(n + 1)p$, if $(n + 1)p$ is not an integer
= $(n + 1)\, p$ and $(n + 1)\, p - 1$, if $(n + 1)\, p$ is an integer.

5.2.4 Recurrence Relation for the Binomial Distribution

For the binomial distribution,

$$\begin{aligned} P(X=x) &= {}^nC_x\, p^x\, q^{n-x} \\ P(X=x+1) &= {}^nC_{x+1}\, p^{x+1}\, q^{n-x-1} \\ \frac{P(X=x+1)}{P(X=x)} &= \frac{{}^nC_{x+1}\, p^{x+1}\, q^{n-x-1}}{{}^nC_x\, p^x\, q^{n-x}} \\ &= \frac{n!}{(x+1)!\,(n-x-1)!} \times \frac{x!\,(n-x)!}{n!} \cdot \frac{p}{q} \\ &= \frac{(n-x)\,(n-x-1)!\,x!}{(x+1)\,x!\,(n-x-1)!} \cdot \frac{p}{q} \\ &= \frac{n-x}{x+1} \cdot \frac{p}{q} \end{aligned}$$

$$P(X=x+1) = \frac{n-x}{x+1} \cdot \frac{p}{q} \cdot P(X=x)$$

5.2.5 Binomial Frequency Distribution

If n independent trials constitute one experiment and this experiment is repeated N times, the frequency of x successes is $N\,P(X=x)$, i.e., $N\ {}^nC_x\ p^x\ q^{n-x}$. This is called expected or theoretical frequency $f(x)$ of a success.

$$\sum_{x=0}^{n} f(x) = N \sum_{x=0}^{n} P(X=x) = N \qquad \left[\because\ \sum_{x=0}^{n} P(X=x) = 1\right]$$

The expected or theoretical frequencies $f(0), f(1), f(2), \ldots, f(n)$ of $0, 1, 2, \ldots, n$, successes are respectively the first, second, third, ..., $(n+1)^{\text{th}}$ term in the expansion of $N(q+p)^n$. The possible number of successes and their frequencies is called a binomial frequency distribution. In practice, the expected frequencies differ from observed frequencies due to chance factor.

Example 1

The mean and standard deviation of a binomial distribution are 5 and 2. Determine the distribution.

Solution

$$\begin{aligned} \mu &= np = 5 \\ \text{SD} &= \sqrt{npq} = 2 \\ npq &= 4 \end{aligned}$$

$$\begin{aligned} \frac{npq}{np} &= \frac{4}{5} \\ \therefore\ q &= \frac{4}{5} \\ p &= 1-q = 1-\frac{4}{5} = \frac{1}{5} \\ np &= 5 \\ n\left(\frac{1}{5}\right) &= 5 \\ \therefore\ n &= 25 \end{aligned}$$

Hence, the binomial distribution is

$$\begin{aligned} P(X=x) &= {}^nC_x\, p^x\, q^{n-x} \\ &= {}^{25}C_x \left(\frac{1}{5}\right)^x \left(\frac{4}{5}\right)^{25-x}, \quad x = 0, 1, 2, \ldots, 25 \end{aligned}$$

Example 2

The mean and variance of a binomial variate are 8 and 6. Find $P(X \geq 2)$.

Solution

$$\begin{aligned} \mu &= np = 8 \\ \sigma^2 &= npq = 6 \\ \frac{npq}{np} &= \frac{6}{8} = \frac{3}{4} \\ \therefore\ q &= \frac{3}{4} \\ p &= 1-q = 1-\frac{3}{4} = \frac{1}{4} \\ np &= 8 \\ n\left(\frac{1}{4}\right) &= 8 \\ \therefore\ n &= 32 \end{aligned}$$

$$\begin{aligned} P(X=x) &= {}^nC_x\, p^x\, q^{n-x} \\ &= {}^{32}C_x \left(\frac{1}{4}\right)^x \left(\frac{3}{4}\right)^{32-x}, \quad x = 0, 1, 2, \ldots, 32 \end{aligned}$$

$$
\begin{aligned}
P(X \ge 2) &= 1 - P(X < 2)\\
&= 1 - [P(X=0) + P(X=1)]\\
&= 1 - \sum_{x=0}^{1} P(X=x)\\
&= 1 - \sum_{x=0}^{1} {}^{32}C_x \left(\frac{1}{4}\right)^x \left(\frac{3}{4}\right)^{32-x}\\
&= 0.9988
\end{aligned}
$$

Example 3

Suppose $P(X=0) = 1 - P(X=1)$. If $E(X) = 3\ \text{Var}(X)$, find $P(X=0)$.

Solution

$$
\begin{aligned}
E(X) &= 3\,\text{Var}(X)\\
np &= 3npq\\
1 &= 3q\\
\therefore\quad q &= \frac{1}{3}\\
p &= 1 - q = 1 - \frac{1}{3} = \frac{2}{3}
\end{aligned}
$$

Let $P(X=1) = p$

$$
\begin{aligned}
P(X=0) &= 1 - P(X=1)\\
&= 1 - p\\
&= 1 - \frac{2}{3}\\
&= \frac{1}{3}
\end{aligned}
$$

Example 4

The mean and variance of a binomial distribution are 4 and $\frac{4}{3}$ respectively. Find $P(X \ge 1)$.

Solution

$$
\begin{aligned}
\mu &= np = 4\\
\sigma^2 &= npq = \frac{4}{3}
\end{aligned}
$$

$$
\begin{aligned}
\frac{npq}{np} &= \frac{\frac{4}{3}}{4} = \frac{1}{3}\\
\therefore\quad q &= \frac{1}{3}\\
p &= 1 - q = 1 - \frac{1}{3} = \frac{2}{3}\\
np &= 4\\
n\left(\frac{2}{3}\right) &= 4\\
\therefore\quad n &= 6
\end{aligned}
$$

$$
\begin{aligned}
P(X=x) &= {}^nC_x\, p^x\, q^{n-x}\\
&= {}^6C_x \left(\frac{2}{3}\right)^x \left(\frac{1}{3}\right)^{6-x}, \quad x = 0, 1, 2, \ldots, 6\\
P(X \ge 1) &= 1 - P(X < 1)\\
&= 1 - P(X=0)\\
&= 1 - {}^6C_0 \left(\frac{2}{3}\right)^0 \left(\frac{1}{3}\right)^6\\
&= 0.9986
\end{aligned}
$$

Example 5

A discrete random variable X has mean 6 and variance 2. If it is assumed that the distribution is binomial, find the probability that $5 \le X \le 7$.

Solution

$$
\begin{aligned}
\mu &= np = 6\\
\sigma^2 &= npq = 2\\
\frac{npq}{np} &= \frac{2}{6} = \frac{1}{3}\\
\therefore\quad q &= \frac{1}{3}\\
p &= 1 - q = 1 - \frac{1}{3} = \frac{2}{3}\\
np &= 6\\
n\left(\frac{2}{3}\right) &= 6\\
\therefore\quad n &= 9
\end{aligned}
$$

$$\begin{aligned} P(X=x) &= {}^nC_x\, p^x\, q^{n-x} \\ &= {}^9C_x \left(\frac{2}{3}\right)^x \left(\frac{1}{3}\right)^{9-x}, \quad x=0,1,2,\dots,9 \end{aligned}$$

$$\begin{aligned} P(5 \le X \le 7) &= P(X=5)+P(X=6)+P(X=7) \\ &= \sum_{x=5}^{7} P(X=x) \\ &= \sum_{x=5}^{7} {}^9C_x \left(\frac{2}{3}\right)^x \left(\frac{1}{3}\right)^{9-x} \\ &= \frac{4672}{6561} \\ &= 0.7121 \end{aligned}$$

Example 6

With the usual notation, find p for a binomial distribution if n = 6 and 9P(X = 4) = P(X = 2).

Solution

For the binomial distribution,

$$P(X=x) = {}^nC_x\, p^x\, q^{n-x}, \quad x=0,1,2,\dots,n$$

$$\begin{aligned} n &= 6 \\ 9P(X=4) &= P(X=2) \\ 9\,{}^6C_4\, p^4\, q^2 &= {}^6C_2\, p^2 q^4 \\ 9p^2 &= q^2 = (1-p)^2 \\ 9p^2 &= 1-2p+p^2 \\ 8p^2+2p-1 &= 0 \\ p &= \frac{-2 \pm \sqrt{4+32}}{2 \times 8} = \frac{-2 \pm 6}{16} = -\frac{1}{2}, \frac{1}{4} \end{aligned}$$

Since probability cannot be negative, $p = \frac{1}{4}$.

Example 7

In a binomial distribution consisting of 5 independent trials, the probability of 1 and 2 successes are 0.4096 and 0.2048 respectively. Find the parameter p of the distribution.

Solution

$$n=5, \quad P(X=1)=0.4096, \quad P(X=2)=0.2048$$

Probability of getting x successes out of 5 trials

$$P(X=x) = {}^nC_x\, p^x\, q^{n-x} = {}^5C_x\, p^x\, q^{5-x}, x=0,1,2,\dots,5$$

$$P(X=1) = {}^5C_1\, pq^4 = 0.4096 \qquad \text{...(1)}$$

$$P(X=2) = {}^5C_2\, p^2\, q^3 = 0.2048 \qquad \text{...(2)}$$

Dividing Eq. (2) by Eq. (1),

$$\begin{aligned} \frac{{}^5C_2\, p^2\, q^3}{{}^5C_1\, p\, q^4} &= \frac{0.2048}{0.4096} \\ \frac{10p}{5q} &= \frac{1}{2} \\ \frac{p}{q} &= \frac{1}{4} \\ 4p &= q = 1-p \\ 5p &= 1 \\ p &= \frac{1}{5} \end{aligned}$$

Example 8

In a binomial distribution, the sum and product of the mean and variance are $\frac{25}{3}$ and $\frac{50}{3}$ respectively. Determine the distribution.

Solution

For the binomial distribution,

$$np + npq = \frac{25}{3} \qquad \text{...(1)}$$

$$np(1+q) = \frac{25}{3}$$

and

$$np(npq) = \frac{50}{3} \qquad \text{...(2)}$$

$$n^2\, p^2\, q = \frac{50}{3}$$

Scanned with CamScanner

Squaring Eq. (1) and then dividing by Eq. (2),

$$\frac{n^2 p^2 (1+q)^2}{n^2 p^2 q} = \frac{\frac{625}{9}}{\frac{50}{3}}$$

$$\frac{1+2q+q^2}{q} = \frac{25}{6}$$

$$6(q^2+2q+1) = 25q$$

$$6q^2 - 13q + 6 = 0$$

$$(2q-3)(3q-2) = 0$$

$$q = \frac{3}{2} \text{ or } q = \frac{2}{3}$$

Since q can not be greater than 1,

$$q = \frac{2}{3}$$

$$p = 1 - q = 1 - \frac{2}{3} = \frac{1}{3}$$

From Eq. (1),

$$n\left(\frac{1}{3}\right)\left(1+\frac{2}{3}\right) = \frac{25}{3}$$

$$\therefore \qquad n = 15$$

Hence, the binomial distribution is

$$P(X=x) = {}^{15}C_x \left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{15-x}, \qquad x = 0, 1, 2, \ldots 15$$

Example 9

If the probability of a defective bolt is $\frac{1}{8}$, *find the (i) mean, and (ii) variance for the distribution of* 640 *defective bolts.*

Solution

$$p = \frac{1}{8}, \quad n = 640$$

$$\mu = np = \frac{640}{8} = 80$$

$$q = 1 - p = 1 - \frac{1}{8} = \frac{7}{8}$$

Variance of the distribution $= npq = 640\left(\frac{1}{8}\right)\left(\frac{7}{8}\right) = 70$

Example 10

In eight throws of a die, 5 or 6 is considered as a success. Find the mean number of success and the standard deviation.

Solution

Let p be the probability of success.

$$p = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

$$q = 1 - p = 1 - \frac{1}{3} = \frac{2}{3}$$

$$n = 8$$

$$\mu = np = 8\left(\frac{1}{3}\right) = \frac{8}{3}$$

$$\text{SD} = \sqrt{npq} = \sqrt{8\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)} = \frac{4}{3}$$

Example 11

4 coins are tossed simultaneously. What is the probability of getting (i) 2 heads? (ii) at least 2 heads? (iii) at most 2 heads?

Solution

Let p be the probability of getting a head in the toss of a coin.

$$p = \frac{1}{2}, \quad q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2}, \quad n = 4$$

The probability of getting x heads when 4 coins are tossed

$$P(X=x) = {}^{n}C_x \, p^x q^{n-x} = {}^{4}C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{4-x}, \quad x = 0, 1, 2, 3, 4$$

(i) Probability of getting 2 heads when 4 coins are tossed

$$P(X=2) = {}^{4}C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2 = \frac{3}{8}$$

(ii) Probability of getting at least two heads when 4 coins are tossed

$$\begin{aligned} P(X \geq 2) &= P(X=2)+P(X=3)+P(X=4) \\ &= \sum_{x=2}^{4} P(X=x) \\ &= \sum_{x=2}^{4} {}^4C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{4-x} \\ &= \frac{11}{16} \end{aligned}$$

(iii) Probability getting at most 2 heads when 4 coins are tossed

$$\begin{aligned} P(X \leq 2) &= P(X=0)+P(X=1)+P(X=2) \\ &= \sum_{x=0}^{2} P(X=x) \\ &= \sum_{x=0}^{2} {}^4C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{4-x} \\ &= \frac{11}{16} \end{aligned}$$

Example 12

Two dice are thrown five times. Find the probability of getting the sum as 7 (i) at least once, (ii) two times, and (iii) $P(1 < X < 15)$.

Solution

In a single throw of two dice, a sum of 7 can occur in 6 ways out of $6 \times 6 = 36$ ways.

$(1, 6), (6, 1), (2, 5), (5, 2), (3, 4), (4, 3)$

Let p be the probability of getting the sum as 7 in a single throw of a pair of dice.

$$p = \frac{6}{36} = \frac{1}{6}, \quad q = 1-p = 1-\frac{1}{6} = \frac{5}{6}, \quad n = 5$$

Probability of getting the sum x times in 5 throws of a pair of dice

$$P(X=x) = {}^nC_x\, p^x\, q^{n-x} = {}^5C_x \left(\frac{1}{6}\right)^x \left(\frac{5}{6}\right)^{5-x}, \quad x = 0, 1, 2, \ldots, 5$$

(i) Probability of getting the sum as 7 at least once in 5 throws of two dice

$$\begin{aligned} P(X \geq 1) &= 1 - P(X=0) \\ &= 1 - {}^5C_0 \left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^5 \end{aligned}$$

$$\begin{aligned} &= 1 - \frac{3125}{7776} \\ &= \frac{4651}{7776} \end{aligned}$$

(ii) Probability of getting the sum as 7 two times in 5 throws of two dice

$$P(X=2) = {}^5C_2 \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^3 = \frac{625}{3888}$$

(iii) Probability of getting the sum as 7 for $P(1 < X < 5)$ in 5 throws of two dice

$$\begin{aligned} P(1 < X < 5) &= P(X=2)+P(X=3)+P(X=4) \\ &= \sum_{x=2}^{4} P(X=x) \\ &= \sum_{x=2}^{4} {}^5C_x \left(\frac{1}{6}\right)^x \left(\frac{5}{6}\right)^{5-x} \\ &= \frac{1525}{7776} \end{aligned}$$

Example 13

If 10% of the screws produced by a machine are defective, find the probability that out of 5 screws chosen at random, (i) none is defective, (ii) one is defective, and (iii) at most two are defective.

Solution

Let p be the probability of defective screws.

$$p = 0.1, \qquad q = 1-p = 1-0.1 = 0.9, \qquad n = 5$$

Probability that x screws out of 5 screws are defective

$$P(X=x) = {}^nC_x\, p^x\, q^{n-x} = {}^5C_x\, (0.1)^x\, (0.9)^{5-x}, \quad x = 0, 1, 2, \ldots, 5$$

(i) Probability that none of the screws out of 5 screws is defective

$$P(X=0) = {}^5C_0\, (0.1)^0\, (0.9)^5 = 0.5905$$

(ii) Probability that one screw out of 5 screws is defective

$$P(X=1) = {}^5C_1\, (0.1)^1\, (0.9)^4 = 0.3281$$

(iii) Probability that at most 2 screws out of 5 screws are defective

$$\begin{aligned} P(X \leq 2) &= P(X=0)+P(X=1)+P(X=2) \\ &= \sum_{x=0}^{2} P(X=x) \\ &= \sum_{x=0}^{2} {}^5C_x (0.1)^x (0.9)^{5-x} \\ &= 0.9914 \end{aligned}$$

Example 14

A multiple-choice test consists of 8 questions with 3 answers to each question (of which only one is correct). A student answers each question by rolling a balanced die and checking the first answer if he gets 1 or 2, the second answer if he gets 3 or 4, and the third answer if he gets 5 or 6. To get a distinction, the student must secure at least 75% correct answers. If there is no negative making, what is the probability that the student secures a distinction? **[Summer 2015]**

Solution

Let p be the probability of getting an answer to a question correctly. There are three answers to each question, out of which only one is correct.

$$p=\frac{1}{3},\quad q=1-p=1-\frac{1}{3}=\frac{2}{3},\quad n=8$$

Probability of getting x correct answers in an 8 questions test

$$P(X=x)={}^nC_x\, p^x\, q^{n-x}={}^8C_x\left(\frac{1}{3}\right)^x\left(\frac{2}{3}\right)^{8-x},\quad x=0,1,2,\ldots,8$$

Probability of securing a distinction, i.e., getting at least 6 correct answers out of the 8 questions

$$\begin{aligned}P(X\le 6)&=P(X=6)+P(X=7)+P(X=8)\\&=\sum_{x=6}^{8}P(X=x)\\&=\sum_{x=6}^{8}{}^8C_x\left(\frac{1}{3}\right)^x\left(\frac{2}{3}\right)^{8-x}\\&=\frac{43}{2187}\\&=0.0197\end{aligned}$$

Example 15

A and B play a game in which their chances of winning are in the ratio 3:2. Find A's chance of winning at least three games out of the five games played.

Solution

Let p be the probability that A wins the game.

$$p=\frac{3}{3+2}=\frac{3}{5},\quad q=1-p=1-\frac{3}{5}=\frac{2}{5},\quad n=5$$

Probability that A wins x games out of 5 games

$$P(X=x)={}^nC_x\, p^x\, q^{n-x}={}^5C_x\left(\frac{3}{5}\right)^x\left(\frac{2}{5}\right)^{5-x},\quad x=0,1,2,\ldots,5$$

Probability that A wins at least 3 games

$$\begin{aligned}P(X\ge 3)&=P(X=3)+P(X=4)+P(X=5)\\&=\sum_{x=3}^{5}P(X=x)\\&=\sum_{x=3}^{5}{}^5C_x\left(\frac{3}{5}\right)^x\left(\frac{2}{5}\right)^{5-x}\\&=\frac{2133}{3125}\\&=0.6826\end{aligned}$$

Example 16

It has been claimed that in 60% of all solar heat installations the utility bill is reduced by at least one-third. Accordingly, what are the probabilities that the utility bill will be reduced by at least one third in (i) four of five installations? (ii) at least four of five installations?

Solution

Let p be the probability that the utility bill is reduced by one-third in the solar heat installations.

$$p=60\%=0.6,\qquad q=1-p=1-0.6=0.4,\qquad n=5$$

Probability that the utility bill is reduced by one-third in x installations out of 5 installations

$$P(X=x)={}^nC_x\, p^x\, q^{n-x}={}^5C_x(0.6)^x\,(0.4)^{5-x},\ x=0,1,2,\ldots,5$$

Probability that the utility bill is reduced by one-third in 4 of 5 installations

$$P(X=5)={}^5C_4\,(0.6)^4\,(0.4)^1=\frac{162}{625}$$

Probability that the utility bill is reduced by one-third in at least 4 of 5 installations

$$\begin{aligned}P(X\ge 4)&=P(X=4)+P(X=5)\\&=\sum_{x=4}^{5}P(X=x)\\&=\sum_{x=4}^{5}{}^5C_x\,(0.6)^x(0.4)^{5-x}\\&=\frac{1053}{3125}\\&=0.337\end{aligned}$$

Example 17

The incidence of an occupational disease in an industry is such that the workers have a 20% chance of suffering from it. What is the probability that out of 6 workers chosen at random, four or more will suffer from the disease?

Solution

Let p be the probability of a worker suffering from the disease.

$$p = 0.2, \quad q = 1 - p = 1 - 0.2 = 0.8, \quad n = 6$$

Probability that x workers will suffer from the disease

$$P(X = x) = {}^{n}C_x\, p^x\, q^{n-x} = {}^{6}C_x\, (0.2)^x\, (0.8)^{6-x}, \quad x = 0, 1, 2, \ldots, 6$$

Probability that 4 or more workers will suffer from the disease

$$\begin{aligned} P(X \ge 4) &= P(X = 4) + P(X = 5) + P(X = 6) \\ &= \sum_{x=4}^{6} P(X = x) \\ &= \sum_{x=4}^{6} {}^{6}C_x\, (0.2)^x (0.8)^{6-x} \\ &= \frac{53}{3125} \\ &= 0.017 \end{aligned}$$

Example 18

The probability that a man aged 60 will live up to 70 is 0.65. What is the probability that out of 10 such men now at 60 at least 7 will live up to 70?

Solution

Let p be the probability that a man will live up to 70.

$$p = 0.65, \quad q = 1 - p = 1 - 0.65 = 0.35, \quad n = 10$$

Probability that x men out of 10 will live up to 70

$$P(X = x) = {}^{n}C_x\, p^x\, q^{n-x} = {}^{10}C_x (0.65)^x\, (0.35)^{10-x}, x = 0, 1, 2, \ldots, 10$$

Probability that at least 7 men out of 10 will live up to 70

$$\begin{aligned} P(X \ge 7) &= P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10) \\ &= \sum_{x=7}^{10} P(X = x) \end{aligned}$$

$$\begin{aligned} &= \sum_{x=7}^{10} {}^{10}C_x\, (0.65)^x\, (0.35)^{10-x} \\ &= 0.5138 \end{aligned}$$

Example 19

In a multiple-choice examination, there are 20 questions. Each question has 4 alternative answers following it and the student must select one correct answer. 4 marks are given for a correct answer and 1 mark is deducted for a wrong answer. A student must secure at least 50% of the maximum possible marks to pass the examination. Suppose a student has not studied at all, so that he answers the questions by guessing only. What is the probability that he will pass the examination?

Solution

Since there are 20 questions and each carries with 4 marks, the maximum marks are 80. If the student solves 12 questions correctly and 8 questions wrongly, he gets 48 – 8 = 40 marks required for passing. If he gets more than 12 correct answers, he gets more than 40 marks. Let p be the probability of getting a correct answer.

$$\therefore \quad p = \frac{1}{4}, \; q = 1 - p = 1 - \frac{1}{4} = \frac{3}{4}, \; n = 20$$

Probability of getting x correct answers out of 20 answers

$$P(X = x) = {}^{n}C_x\, p^x\, q^{n-x} = {}^{20}C_x \left(\frac{1}{4}\right)^x \left(\frac{3}{4}\right)^{20-x}, \; x = 0, 1, 2, \ldots, 20$$

Probability of passing the examination, i.e., probability of getting at least 12 correct answers out of 20 answers

$$\begin{aligned} P(X \ge 12) &= \sum_{x=12}^{20} P(X = x) \\ &= \sum_{x=12}^{20} {}^{20}C_x \left(\frac{1}{4}\right)^x \left(\frac{3}{4}\right)^{20-x} \\ &= 9.3539 \times 10^{-4} \end{aligned}$$

Example 20

The probability of a man hitting a target is $\frac{1}{3}$. (i) If he fires 5 times, what is the probability of his hitting the target at least twice? (ii) How many times must he fire so that the probability of his hitting the target at least once is more than 90%?

Solution

Let p be probability of hitting a target.

$$p=\frac{1}{3},\quad q=1-p=1-\frac{1}{3}=\frac{2}{3},\quad n=5$$

Probability of hitting the target x times out of 5 times

$$P(X=x)={}^nC_x\,p^x\,q^{n-x}={}^5C_x\left(\frac{1}{3}\right)^x\left(\frac{2}{3}\right)^{5-x},\quad x=0,1,2,\ldots,5$$

(i) Probability of hitting the target at least twice out of 5 times

$$P(X\ge 2)=P(X=2)+P(X=3)+P(X=4)+P(X=5)$$
$$=\sum_{x=2}^{5}P(X=x)$$
$$=\sum_{x=2}^{5}{}^5C_x\left(\frac{1}{3}\right)^x\left(\frac{2}{3}\right)^{5-x}$$
$$=\frac{131}{243}$$
$$=0.5391$$

(ii) Probability of hitting the target at least once out of 5 times

$$P(X\ge 1)>0.9$$
$$1-P(X=0)>0.9$$
$$1-{}^nC_0\left(\frac{1}{3}\right)^0\left(\frac{2}{3}\right)^n>0.9$$
$$1-\left(\frac{2}{3}\right)^n>0.9$$

For $n=6$, $1-\left(\frac{2}{3}\right)^6=0.9122$

Hence, the man must fire 6 times so that the probability of hitting the target at lest once is more than 90%.

Example 21

In sampling a large number of parts manufactured by a machine, the mean number of defectives in a sample of 20 is 2. Out of 1000 such samples, how many would be expected to contain exactly two defective parts? **[Summer 2015]**

Solution

Let p be the probability of parts being defective.

$$\mu=np=2,\quad n=20,\quad N=1000$$
$$np=2$$
$$20(p)=2$$
$$\therefore\ p=0.1$$
$$q=1-p=1-0.1=0.9$$

Probability that the samples contain x defective parts out of 20 parts

$$P(X=x)={}^nC_x p^x\,q^{n-x}={}^{20}C_x(0.1)^x(0.9)^{20-x},\ x=0,1,2,\ldots,20$$

Probability that the samples contain exactly 2 defective parts

$$P(X=2)={}^{20}C_2(0.1)^2\,(0.9)^{18}=0.2852$$

Expected number of samples to contain exactly 2 defective parts $=N\,P(X=2)$

$$=1000\,(0.2852)$$
$$=285.2$$
$$\approx 285$$

Example 22

An irregular 6-faced die is thrown such that the probability that it gives 3 even numbers in 5 throws is twice the probability that it gives 2 even numbers in 5 throws. How many sets of exactly 5 trials can be expected to give no even number out of 2500 sets?

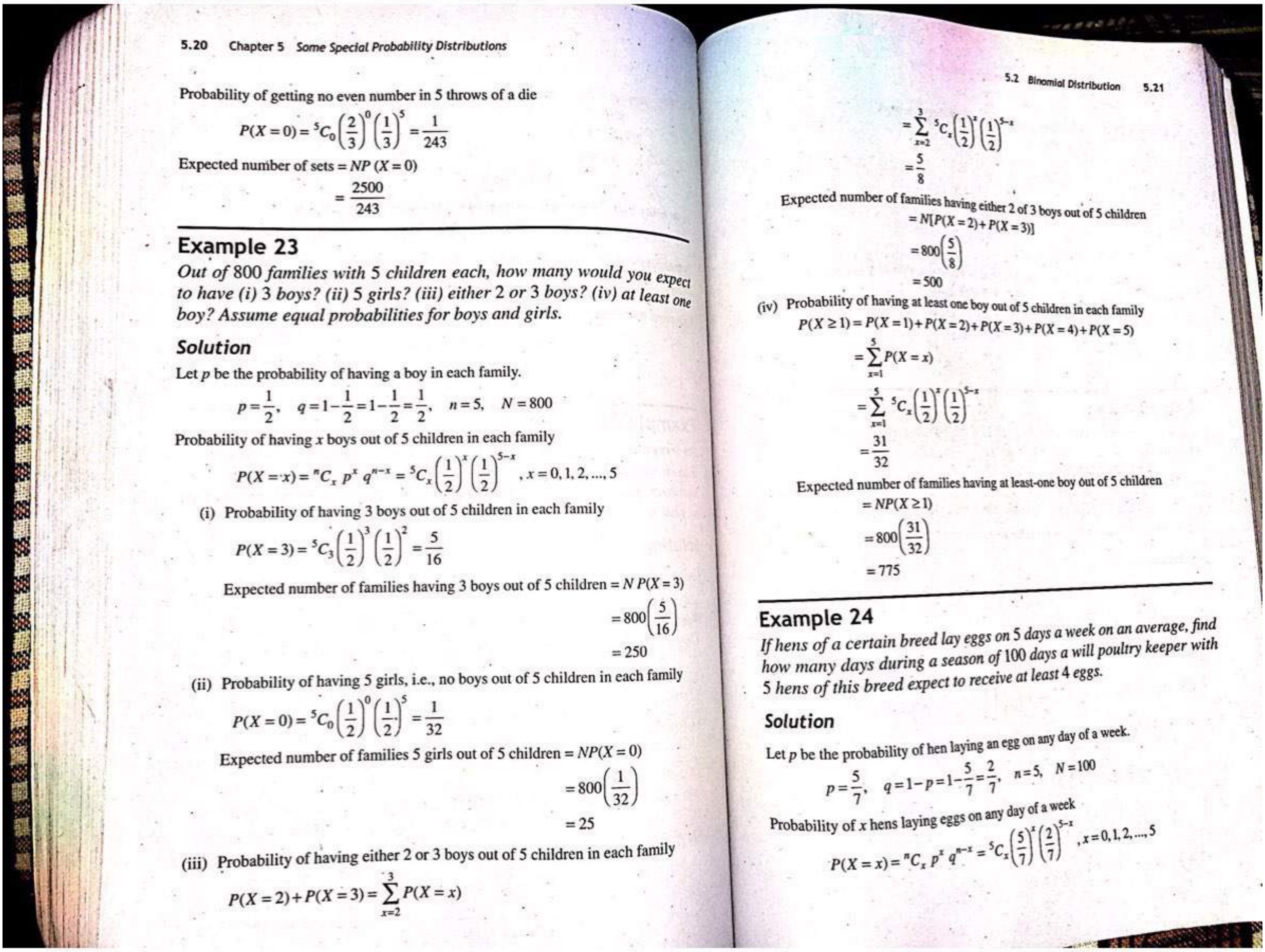
Solution

Let p be the probability of getting an even number in a throw of a die.

$$n=5,\quad N=2500$$

Probability of getting x even numbers in 5 throws of a die

$$P(X=x)={}^nC_x\,p^x\,q^{n-x}={}^5C_x\,p^xq^{5-x},\ x=0,1,2,\ldots,5$$
$$P(X=3)=2\,P(X=2)$$
$${}^5C_3\,p^3\,q^2=2\left({}^5C_2\,p^2\,q^3\right)$$
$$10\,p^3\,q^2=20\,p^2\,q^3$$
$$p=2q$$
$$p=2(1-p)=2-2p$$
$$\therefore\ p=\frac{2}{3}$$
$$q=1-p=1-\frac{2}{3}=\frac{1}{3}$$



Probability of getting no even number in 5 throws of a die

$$P(X=0)={}^5C_0\left(\frac{2}{3}\right)^0\left(\frac{1}{3}\right)^5=\frac{1}{243}$$

Expected number of sets $= NP\ (X=0)$

$$=\frac{2500}{243}$$

Example 23

Out of 800 families with 5 children each, how many would you expect to have (i) 3 boys? (ii) 5 girls? (iii) either 2 or 3 boys? (iv) at least one boy? Assume equal probabilities for boys and girls.

Solution

Let p be the probability of having a boy in each family.

$$p=\frac{1}{2},\quad q=1-\frac{1}{2}=1-\frac{1}{2}=\frac{1}{2},\quad n=5,\quad N=800$$

Probability of having x boys out of 5 children in each family

$$P(X=x)={}^nC_x\,p^x\,q^{n-x}={}^5C_x\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{5-x},\ x=0,1,2,\ldots,5$$

(i) Probability of having 3 boys out of 5 children in each family

$$P(X=3)={}^5C_3\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^2=\frac{5}{16}$$

Expected number of families having 3 boys out of 5 children $= N\,P(X=3)$

$$=800\left(\frac{5}{16}\right)$$

$$=250$$

(ii) Probability of having 5 girls, i.e., no boys out of 5 children in each family

$$P(X=0)={}^5C_0\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^5=\frac{1}{32}$$

Expected number of families 5 girls out of 5 children $= NP(X=0)$

$$=800\left(\frac{1}{32}\right)$$

$$=25$$

(iii) Probability of having either 2 or 3 boys out of 5 children in each family

$$P(X=2)+P(X=3)=\sum_{x=2}^{3}P(X=x)$$

$$=\sum_{x=2}^{3}{}^5C_x\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{5-x}$$

$$=\frac{5}{8}$$

Expected number of families having either 2 of 3 boys out of 5 children

$$=N[P(X=2)+P(X=3)]$$

$$=800\left(\frac{5}{8}\right)$$

$$=500$$

(iv) Probability of having at least one boy out of 5 children in each family

$$P(X\ge 1)=P(X=1)+P(X=2)+P(X=3)+P(X=4)+P(X=5)$$

$$=\sum_{x=1}^{5}P(X=x)$$

$$=\sum_{x=1}^{5}{}^5C_x\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{5-x}$$

$$=\frac{31}{32}$$

Expected number of families having at least-one boy out of 5 children

$$=NP(X\ge 1)$$

$$=800\left(\frac{31}{32}\right)$$

$$=775$$

Example 24

If hens of a certain breed lay eggs on 5 days a week on an average, find how many days during a season of 100 days a will poultry keeper with 5 hens of this breed expect to receive at least 4 eggs.

Solution

Let p be the probability of hen laying an egg on any day of a week.

$$p=\frac{5}{7},\quad q=1-p=1-\frac{5}{7}=\frac{2}{7},\quad n=5,\quad N=100$$

Probability of x hens laying eggs on any day of a week

$$P(X=x)={}^nC_x\,p^x\,q^{n-x}={}^5C_x\left(\frac{5}{7}\right)^x\left(\frac{2}{7}\right)^{5-x},\ x=0,1,2,\ldots,5$$

Scanned with CamScanner

Probability of receiving at least 4 eggs on any day of a week

$$
\begin{aligned}
P(X \geq 4) &= P(X=4)+P(X=5) \\
&= \sum_{x=4}^{5} P(X=x) \\
&= \sum_{x=4}^{5} {}^{5}C_x \left(\frac{5}{7}\right)^x \left(\frac{2}{7}\right)^{5-x} \\
&= 0.5578
\end{aligned}
$$

Expected number of days during a season of 100 days, a poultry keeper with 5 hens of this breed will receive at least 4 eggs $= N\,P(X \geq 4)$

$$
\begin{aligned}
&= 100\,(0.5578) \\
&= 55.78 \\
&\approx 56
\end{aligned}
$$

Example 25

Seven unbiased coins are tossed 128 *times and the number of heads obtained is noted as given below:*

No. of heads	0	1	2	3	4	5	6	7
Frequency	7	6	19	35	30	23	7	1

Fit a binomial distribution to the data.

Solution

Since the coin is unbiased,

$$p=\frac{1}{2}, \quad q=\frac{1}{2}, n=7, N=128$$

For binomial distribution,

$$P(X=x)={}^{n}C_x\, p^x\, q^{n-x}={}^{7}C_x\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{7-x}, x=0,1,2,\ldots,7$$

Theoretical or expected frequency $f(x)=N\,P(X=x)$

$$f(x)=128\;{}^{7}C_x\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{7-x}=128\;{}^{7}C_x\left(\frac{1}{2}\right)^7$$

$$f(0)=128\;{}^{7}C_0\left(\frac{1}{2}\right)^7=1$$

$$f(1)=128\;{}^{7}C_1\left(\frac{1}{2}\right)^7=7$$

$$f(2)=128\;{}^{7}C_2\left(\frac{1}{2}\right)^7=21$$

$$f(3)=128\;{}^{7}C_3\left(\frac{1}{2}\right)^7=35$$

$$f(4)=128\;{}^{7}C_4\left(\frac{1}{2}\right)^7=35$$

$$f(5)=128\;{}^{7}C_5\left(\frac{1}{2}\right)^7=21$$

$$f(6)=128\;{}^{7}C_6\left(\frac{1}{2}\right)^7=7$$

$$f(7)=128\;{}^{7}C_7\left(\frac{1}{2}\right)^7=1$$

Binomial distribution

No. of heads x	0	1	2	3	4	5	6	7
Expected binomial frequency $f(x)$	1	7	21	35	35	21	7	1

Example 26

Fit a binomial distribution to the following data:

x	0	1	2	3	4	5
f	2	14	20	34	22	8

Solution

$$
\begin{aligned}
\text{Mean} &= \frac{\sum fx}{\sum f} \\
&= \frac{2(0)+14(1)+20(2)+34(3)+22(4)+8(5)}{2+14+20+34+22+8} \\
&= \frac{284}{100} \\
&= 2.84
\end{aligned}
$$

For binomial distribution,

$$n=5$$

$$\mu = np = 2.84$$
$$5p = 2.84$$
$$\therefore \; p = 0.568$$
$$q = 1 - p = 1 - 0.568 = 0.432$$
$$P(X = x) = {}^nC_x \, p^x \, q^{n-x} = {}^5C_x (0.568)^x (0.432)^{5-x}, \quad x = 0, 1, 2, \ldots, 5$$
$$N = \sum f = 100$$

Theoretical or expected frequency $f(x) = N\,P(X = x)$

$$f(x) = 100\;{}^5C_x\,(0.568)^x (0.432)^{5-x}$$
$$f(0) = 100\;{}^5C_0\,(0.568)^0\,(0.432)^5 = 1.505 \approx 1.5$$
$$f(1) = 100\;{}^5C_1\,(0.568)^1\,(0.432)^4 = 9.89 \approx 10$$
$$f(2) = 100\;{}^5C_2\,(0.568)^2\,(0.432)^3 = 26.01 \approx 26$$
$$f(3) = 100\;{}^5C_2\,(0.568)^3\,(0.432)^2 = 34.2 \approx 34$$
$$f(4) = 100\;{}^5C_2\,(0.568)^4\,(0.432)^1 = 22.48 \approx 22$$
$$f(5) = 100\;{}^5C_2\,(0.568)^5\,(0.432)^0 = 5.91 \approx 6$$

Binomial Distribution

x	0	1	2	3	4	5
Expected binomial frequency	1.5	10	26	34	22	6

EXERCISE 5.1

1. Find the fallacy if any in the following statements:
   (a) The mean of a binomial distribution is 6 and SD is 4.
   (b) The mean of a binomial distribution is 9 and its SD is 4.

[**Ans.:** (a) False, $q = \frac{8}{3}$ is impossible
(b) False, $q = \frac{19}{9}$ is impossible]

2. The mean and variance of a binomial distribution are 3 and 1.2 respectively. Find n, p, and $P(X < 4)$.

[**Ans.:** 5, 0.6, $\frac{2068}{3125}$]

3. Find the binomial distribution if the mean is 5 and the variance is $\frac{10}{3}$. Find $P(X = 2)$.

[**Ans.:** $P(X = x) = {}^{25}C_x \left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{25-x}$, 0.003]

4. In a binomial distribution, the mean and variance are 4 and 3 respectively. Find $P(X \geq 1)$.

[**Ans.:** 0.9899]

5. The odds in favour of X winning a game against Y are 4:3. Find the probability of Y winning 3 games out of 7 played.

[**Ans.:** 0.0929]

6. On an average, 3 out of 10 students fail in an examination. What is the probability that out of 10 students that appear for the examination none will fail?

[**Ans.:** 0.0282]

7. If on the average rain falls on 10 days in every thirty, find the probability (i) that the first three days of a week will be fine and remaining wet, and (ii) that rain will fall on just three days of a week.

[**Ans.:** (i) $\frac{8}{2187}$ (ii) $\frac{280}{2187}$]

8. Two unbiased dice are thrown three times. Find the probability that the sum nine would be obtained (i) once, and (ii) twice.

[**Ans.:** (i) 0.26 (ii) 0.03]

9. For special security in a certain protected area, it was decided to put three lightbulbs on each pole. If each bulb has probability p of burning out in the first 100 hours of service, calculate the probability that at least one of them is still good after 100 hours. If $p = 0.3$, how many bulbs would be needed on each pole to ensure with 99% safety that at least one is good after 100 hours?

[**Ans.:** (i) $1 - p^3$ (ii) 4]

10. It is known from past records that 80% of the students in a school do their homework. Find the probability that during a random check of 10 students, (i) all have done their homework, (ii) at the most two have not done their homework, and (iii) at least one has not done the homework.

[**Ans.:** (i) 0.1074 (ii) 0.6778 (iii) 0.8926]

11. An insurance salesman sells policies to 5 men, all of identical age and good health. According to the actuarial tables, the probability that a man of this particular age will be alive 30 years hence is $\frac{2}{3}$. Find the probability that 30 years hence (i) at least 1 man will be alive, (ii) at least 3 men will be alive, and (iii) all 5 men will be alive.

[Ans.: (i) $\frac{242}{243}$ (ii) $\frac{64}{81}$ (iii) $\frac{32}{243}$]

12. A company has appointed 10 new secretaries out of which 7 are trained. If a particular executive is to get three secretaries selected at random, what is the chance that at least one of them will be untrained?

[Ans.: 0.7083]

13. The overall pass rate in a university examination is 70%. Four candidates take up such an examination. What is the probability that (i) at least one of them will pass? (ii) all of them will pass the examination?

[Ans.: (i) 0.9919 (ii) 0.7599]

14. The normal rate of infection of a certain disease in animals is known to be 25%. In an experiment with a new vaccine, it was observed that none of the animals caught the infection. Calculate the probability of the observed result.

[Ans.: $\frac{729}{4096}$]

15. Suppose that weather records show that on the average, 5 out of 31 days in October are rainy days. Assuming a binomial distribution with each day of October as an independent trial, find the probability that the next October will have at most three rainy days.

[Ans.: 0.2403]

16. Assuming that half the population of a village is female and assuming that 100 samples each of 10 individuals are taken, how many samples would you expect to have 3 or less females?

[Ans.: 17]

17. Assuming that half the population of a town is vegetarian so that the chance of an individual being vegetarian is $\frac{1}{2}$, and assuming that 100 investigators can take a sample of 10 individuals to see whether they are vegetarians, how many investigators would you expect to report that three people or less in the sample were vegetarians?

[Ans.: 17]

18. The probability of failure in a physics practical examination is 20%. If 25 batches of 6 students each take the examination, in how many batches of 4 or more students would pass?

[Ans.: 23]

19. A lot contains 1% defective items. What should be the number of items in a lot so that the probability of finding at least one defective item in it is at least 0.95?

[Ans.: 299]

20. The probability that a bomb will hit the target is 0.2. Two bombs are required to destroy the target. If six bombs are used, find the probability that the target will be destroyed.

[Ans.: 0.3447]

21. Out of 1000 families with 4 children each, how many would you expect to have (i) 2 boys and 2 girls? (ii) at least one boy? (iii) no girl? (iv) at most 2 girls?

[Ans.: (i) 375 (ii) 938 (iii) 63 (iv) 69]

22. In a sampling of a large number of parts produced by a machine, the mean number of defectives in a sample of 20 is 2. Out of 1000 such samples, how many samples would you expect to contain at least 3 defectives?

[Ans.: 323]

23. Five pair coins are tossed 3200 times, find the frequency distribution of the number of heads obtained. Also, find the mean and SD.

[Ans.: (i) 100, 500, 1000, 1000, 500, 100 (ii) 1600 (iii) 28.28]

24. Fit a binomial distribution to the following data:

x	0	1	2	3	4
f	12	66	109	59	10

[Ans.: 17, 67, 96, 61, 15]

5.3 POISSON DISTRIBUTION

A random variable X is said to follow poisson distribution if the probability of x is given by

$$P(X=x)=p(x)=\frac{e^{-\lambda}\lambda^{x}}{x!},\quad x=0,1,2,\dots$$

where λ is called the *parameter of the distribution*.

5.3.1 Poisson Approximation to the Binomial Distribution

Poisson distribution is a limiting case of binomial distribution under the following conditions:

(i) The number of trials should be infinitely large, i.e., $n \to \infty$.
(ii) The probability of successes p for each trial should be very small, i.e., $p \to 0$.
(iii) $np = \lambda$ should be finite where λ is a constant.

The binomial distribution is

$$\begin{aligned} P(X=x) &= {}^nC_x\, p^x q^{n-x} \\ &= {}^nC_x \left(\frac{p}{q}\right)^x q^n \\ &= {}^nC_x \left(\frac{p}{1-p}\right)^x (1-p)^n \end{aligned}$$

Putting $p = \dfrac{\lambda}{n}$,

$$\begin{aligned} P(X=x) &= \frac{n(n-1)(n-2)\cdots(n-x+1)}{x!}\left(\frac{\frac{\lambda}{n}}{1-\frac{\lambda}{n}}\right)^x \left(1-\frac{\lambda}{x}\right)^n \\ &= \frac{n(n-1)(n-2)\cdots(n-x+1)}{x!}\,\frac{\lambda^x}{n^x}\,\frac{1}{\left(1-\frac{\lambda}{n}\right)^x}\left(1-\frac{\lambda}{x}\right)^n \\ &= \frac{n(n-1)(n-2)\cdots(n-x+1)}{x!}\,\frac{\lambda^x}{n^x}\left(1-\frac{\lambda}{n}\right)^{n-x} \\ &= \frac{1\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left[1-\left(\frac{x-1}{n}\right)\right]}{x!}\,\lambda^x\left(1-\frac{\lambda}{n}\right)^{n-x} \end{aligned}$$

Since $\displaystyle\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n-x} = e^{-\lambda}$

and $\displaystyle\lim_{n\to\infty}\left(1-\frac{1}{n}\right) = \lim_{n\to\infty}\left(1-\frac{2}{n}\right) = 1$

Taking the limits of both the sides as $n \to \infty$,

$$P(X=x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \ldots, \infty$$

5.3.2 Examples of Poisson Distribution

(i) Number of defective bulbs produced by a reputed company
(ii) Number of telephone calls per minute at a switchboard
(iii) Number of cars passing a certain point in one minute
(iv) Number of printing mistakes per page in a large text
(v) Number of persons born blind per year in a large city

5.3.3 Conditions of Poisson Distribution

The Poisson distribution holds under the following conditions:

(i) The random variable X should be discrete.
(ii) The numbers of trials n is very large.
(iii) The probability of success p is very small (very close to zero).
(iv) $\lambda = np$ is finite.
(v) The occurrences are rare.

5.3.4 Constants of the Poisson Distribution

1. Mean of the Poisson Distribution

$$\begin{aligned} E(X) &= \sum_{x=0}^{\infty} x\, p(x) \\ &= \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} \\ &= \sum_{x=0}^{\infty} \frac{x e^{-\lambda}\lambda\lambda^{x-1}}{x!} \\ &= e^{-\lambda}\cdot\lambda\sum_{x=1}^{\infty}\frac{x\lambda^{x-1}}{x!} \\ &= \lambda e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!} \qquad \left[\because \frac{x}{x!} = \frac{1}{(x-1)!}\right] \\ &= \lambda e^{-\lambda}\left(1+\lambda+\frac{\lambda^2}{2!}+\cdots\right) \\ &= \lambda e^{-\lambda} e^{\lambda} \\ &= \lambda \end{aligned}$$

2. Variance of the Poisson Distribution

$$
\begin{aligned}
\text{Var}(X) &= E(X^2)-\mu^2\\
&=\sum_{x=0}^{\infty} x^2\, p(x)-\mu^2\\
&=\sum_{x=0}^{\infty} x^2 \frac{e^{-\lambda}\lambda^x}{x!}-\lambda^2\\
&=\sum_{x=0}^{\infty} x[(x-1)+x]\frac{e^{-\lambda}\lambda^x}{x!}-\lambda^2\\
&=\sum_{x=0}^{\infty}\frac{x(x-1)e^{-\lambda}\lambda^x}{x!}+\sum_{x=0}^{\infty}\frac{x\,e^{-\lambda}\lambda^x}{x!}-\lambda^2\\
&=\sum_{x=0}^{\infty}\frac{x(x-1)e^{-\lambda}\,\lambda^{x-2}\,\lambda^2}{x(x-1)(x-2)\cdots 1}+\lambda-\lambda^2\\
&=e^{-\lambda}\lambda^2\sum_{x=2}^{\infty}\frac{\lambda^{x-2}}{(x-2)!}+\lambda-\lambda^2\\
&=e^{-\lambda}\,\lambda^2\left(1+\lambda+\frac{\lambda^2}{2!}+\cdots\right)+\lambda-\lambda^2\\
&=-e^{\lambda}\,e^{-\lambda}\,\lambda^2+\lambda-\lambda^2\\
&=\lambda^2+\lambda-\lambda^2\\
&=\lambda
\end{aligned}
$$

3. Standard Deviation of the Poisson Distribution

$$\text{SD}=\sqrt{\text{Variance}}=\sqrt{\lambda}$$

4. Mode of the Poisson Distribution

Mode is the value of x for which the probability $p(x)$ is maximum.

$$p(x)\ge p(x+1) \quad\text{and}\quad p(x)\ge p(x-1)$$

When $p(x)\ge p(x+1)$,

$$
\begin{aligned}
\frac{e^{-\lambda}\,\lambda^x}{x!} &\ge \frac{e^{-\lambda}\,\lambda^{x+1}}{(x+1)!}\\
1 &\ge \frac{\lambda}{x+1}\\
(x+1) &\ge \lambda\\
x &\ge \lambda-1 \qquad \text{...(5.1)}
\end{aligned}
$$

Similarly, for $p(x)\ge p(x-1)$,

$$x\le\lambda \qquad \text{...(5.2)}$$

Combining Eqs (5.1) and (5.2),

$$\lambda-1\le x\le\lambda$$

Hence, the mode of the Poisson distribution lies between $\lambda-1$ and λ.

Case I If λ is an integer then $\lambda-1$ is also an integer. The distribution is bimodal and the two modes are $\lambda-1$ and λ.

Case II If λ is not an integer, the distribution is unimodal and the mode of the Poisson distribution is an integral part of λ. The mode is the integer between $\lambda-1$ and λ.

5.3.5 Recurrence Relation for the Poisson Distribution

For the Poisson distribution,

$$
\begin{aligned}
p(x) &= \frac{e^{-\lambda}\,\lambda^x}{x!}\\
p(x+1) &= \frac{e^{-\lambda}\,\lambda^{x+1}}{(x+1)!}\\
\frac{p(x+1)}{p(x)} &= \frac{e^{-\lambda}\,\lambda^{x+1}}{(x+1)!}\cdot\frac{x!}{e^{-\lambda}\,\lambda^x}\\
&=\frac{\lambda}{x+1}\\
p(x+1) &= \frac{\lambda}{x+1}\,p(x)
\end{aligned}
$$

Example 1

Find out the fallacy if any in the statement. "The mean of a Poisson distribution is 2 and the variance is 3."

Solution

In a Poisson distribution, the mean and variance are same. Hence, the above statement is false.

Example 2

If the mean of the Poisson distribution is 4, find

$$P(\lambda-2\sigma<X<\lambda+2\sigma).$$

Solution

For a Poisson distribution,

Variance $=\lambda$

$$\text{Mean} = \lambda = 4, \quad \sigma = 2$$

$$P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-4}4^x}{x!}, \quad x = 0, 1, 2, \ldots$$

$$\begin{aligned} P(\lambda - 2\sigma < X < \lambda + 2\sigma) &= P(0 < X < 8) \\ &= \sum_{x=1}^{7} P(X=x) \\ &= \sum_{x=1}^{7} \frac{e^{-4}4^x}{x!} \\ &= 0.9306 \end{aligned}$$

Example 3

If the mean of a Poisson variable is 1.8, *find (i)* $P(X > 1)$, *(ii)* $P(X = 5)$, and *(iii)* $P(0 < X < 5)$.

Solution

For a Poisson distribution,

$$\lambda = 1.8$$

$$P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-1.8}1.8^x}{x!}, \quad x = 0, 1, 2, \ldots$$

(i) $$\begin{aligned} P(X > 1) &= 1 - P(X \le 1) \\ &= 1 - [P(X=0) + P(X=1)] \\ &= 1 - \sum_{x=0}^{1} P(X=x) \\ &= 1 - \sum_{x=0}^{1} \frac{e^{-1.8}1.8^x}{x!} \\ &= 0.5372 \end{aligned}$$

(ii) $$P(X=5) = \frac{e^{-1.8}1.8^5}{5!} = 0.026$$

(iii) $$\begin{aligned} P(0 < X < 5) &= P(X=1) + P(X=2) + P(X=3) + P(X=4) \\ &= \sum_{x=1}^{4} P(X=x) \\ &= \sum_{x=1}^{4} \frac{e^{-1.8}1.8^x}{x!} \\ &= 0.7983 \end{aligned}$$

Example 4

If a random variable has a Poisson distribution such that $P(X = 1) = P(X = 2)$, *find (i) the mean of the distribution, (ii)* $P(X = 4)$, *(iii)* $P(X \ge 1)$, *and (iv)* $P(1 < X < 4)$.

Solution

For a Poisson distribution,

$$P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$

(i) $$\begin{aligned} P(X=1) &= P(X=2) \\ \frac{e^{-\lambda}\lambda^1}{1!} &= \frac{e^{-\lambda}\lambda^2}{2!} \\ \lambda^2 &= 2\lambda \\ \lambda^2 - 2\lambda &= 0 \\ \lambda(\lambda - 2) &= 0 \\ \lambda = 0 &\text{ or } \lambda = 2 \end{aligned}$$

Since $\lambda \ne 0$, $\quad \lambda = 2$

Hence, $$P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-2}2^x}{x!}, \quad x = 0, 1, 2, \ldots$$

(ii) $$P(X=4) = \frac{e^{-2}2^4}{4!} = 0.9022$$

(iii) $$\begin{aligned} P(X \ge 1) &= 1 - P(X < 1) \\ &= 1 - P(X=0) \\ &= 1 - \frac{e^{-2}2^0}{0!} \\ &= 0.8647 \end{aligned}$$

(iv) $$\begin{aligned} P(1 < X < 4) &= P(X=2) + P(X=3) \\ &= \sum_{x=2}^{3} P(X=x) \\ &= \sum_{x=2}^{3} \frac{e^{-2}2^x}{x!} \\ &= 0.4511 \end{aligned}$$

Example 5

If X is a Poisson variate such that $P(X = 0) = P(X = 1)$, find $P(X = 0)$ and using recurrence relation formula, find the probabilities at $x = 1, 2, 3, 4,$ and 5.

Solution

For a Poisson distribution,

$$P(X=x)=\frac{e^{-\lambda}\lambda^x}{x!}, \quad x=0,1,2,\ldots$$

$$P(X=0)=P(X=1)$$

$$\frac{e^{-\lambda}\lambda^0}{0!}=\frac{e^{-\lambda}\lambda^1}{1!}$$

$$\lambda=1$$

Hence, $$P(X=x)=\frac{e^{-\lambda}1^x}{x!}, \qquad x=0,1,2,\ldots$$

(i) $$P(X=0)=\frac{e^{-\lambda}\lambda^0}{0!}=0.3678$$

(ii) By recurrence relation,

$$p(x+1)=\frac{\lambda}{x+1}p(x)$$

$$p(x+1)=\frac{1}{x+1}p(x) \qquad [\because \lambda=1]$$

$$p(1)=p(0)=0.3678$$

$$p(2)=\frac{1}{2}p(1)=\frac{1}{2}(0.3678)=0.1839$$

$$p(3)=\frac{1}{3}p(2)=\frac{1}{3}(0.1839)=0.0613$$

$$p(4)=\frac{1}{4}p(3)=\frac{1}{4}(0.0613)=0.015325$$

$$p(5)=\frac{1}{5}p(4)=\frac{1}{5}(0.015325)=0.003065$$

Example 6

If the variance of a Poisson variate is 3, find the probability that (i) $X = 0$, (ii) $0 < X \le 3$, and (iii) $1 \le X < 4$.

Solution

For a Poisson distribution,

Variance = Mean = $\lambda = 3$

$$P(X=x)=\frac{e^{-\lambda}\lambda^x}{x!}=\frac{e^{-3}3^x}{x!}, \quad x=0,1,2,\ldots$$

(i) $$P(X=0)=\frac{e^{-3}3^0}{0!}=0.0498$$

(ii) $$P(0<X\le 3)=P(X=1)+P(X=2)+P(X=3)$$

$$=\sum_{x=1}^{3}P(X=x)$$

$$=\sum_{x=1}^{3}\frac{e^{-3}3^x}{x!}$$

$$=0.5974$$

(iii) $$P(1\le X<4)=P(X=1)+P(X=2)+P(X=3)$$

$$=\sum_{x=1}^{3}P(X=x)$$

$$=\sum_{x=1}^{3}\frac{e^{-3}3^x}{x!}$$

$$=0.5974$$

Example 7

If a Poisson distribution is such that $\frac{3}{2}P(X=1)=P(X=3)$, find (i) $P(X\ge 1)$, (ii) $P(X\le 3)$, and (iii) $P(2\le X\le 5)$.

Solution

For a Poisson distribution,

$$P(X=x)=\frac{e^{-\lambda}\lambda^x}{x!}, \quad x=0,1,2,\ldots$$

$$\frac{3}{2}P(X=1)=P(X=3)$$

$$\frac{3}{2}\frac{e^{-\lambda}\lambda^1}{1!}=\frac{e^{-\lambda}\lambda^3}{3!}$$

$$\frac{3}{2}\lambda=\frac{\lambda^3}{6}$$

$$\lambda^3-9\lambda=0$$

$$\lambda(\lambda^2-9)=0$$

$$\lambda=0,3,-3$$

Since $\lambda>0,\quad \lambda=3$

Hence, $P(x=x)=\dfrac{e^{-3}3^x}{x!},\qquad x=0,1,2,\ldots$

(i) $P(X\ge 1)=1-P(X<1)$

$$=1-P(X=0)$$

$$=1-\frac{e^{-3}3^0}{0!}$$

$$=0.9502$$

(ii) $P(X\le 3)=P(X=0)+P(X=1)+P(X=2)+P(X=3)$

$$=\sum_{x=0}^{3}P(X=x)$$

$$=\sum_{x=0}^{3}\frac{e^{-3}3^x}{x!}$$

$$=0.6472$$

(iii) $P(2\le X\le 5)=P(X=2)+P(X=3)+P(X=4)+P(X=5)$

$$=\sum_{x=2}^{5}P(X=x)$$

$$=\sum_{x=2}^{5}\frac{e^{-3}\,3^x}{x!}$$

$$=0.7169$$

Example 8

If X is a Poisson variate such that

$$P(X=2)=9\,P(X=4)+90\,P(X=6)$$

Find (i) the mean of X, (ii) the variance of X, (iii) $P(X<2)$, (iv) $P(X>4)$, and (v) $P(X\ge 1)$.

Solution

For a Poisson distribution,

$$P(X=x)=\frac{e^{-\lambda}\lambda^x}{x!},\qquad x=0,1,2,\ldots$$

$$P(X=2)=9P(X=4)+90P(X=6)$$

$$\frac{e^{-\lambda}\lambda^2}{2!}=9\frac{e^{-\lambda}\lambda^4}{4!}+90\frac{e^{-\lambda}\lambda^6}{6!}$$

$$=e^{-\lambda}\lambda^2\left(\frac{9\lambda^2}{4!}+\frac{90\lambda^4}{6!}\right)$$

$$\frac{1}{2}=\frac{9\lambda^2}{4!}+\frac{90\lambda^4}{6!}$$

$$\frac{1}{2}=\frac{3\lambda^2}{8}+\frac{\lambda^4}{8}$$

$$\lambda^4+3\lambda^2-4=0$$

$$\lambda^2=-\frac{3\pm\sqrt{9+16}}{2}=\frac{-3\pm 5}{2}=1,-4$$

Since $\lambda>0$, $\lambda^2=1$

(i) Mean $=\lambda=1$

(ii) Variance $=\lambda=1$

$$P(X=x)=\frac{e^{-1}1^x}{x!},\qquad x=0,1,2,\ldots$$

(iii) $P(X<2)=P(X=0)+P(X=1)$

$$=\sum_{x=0}^{1}\frac{e^{-1}1^x}{x!}$$

$$=0.7358$$

(iv) $P(X>4)=1-P(X\le 4)$

$$=1-[P(X=0)+(X=1)+P(X=2)+P(X=3)+P(X=4)]$$

$$=1-\sum_{x=0}^{4}\frac{e^{-1}1^x}{x!}$$

$$=0.00366$$

(v) $P(X\ge 1)=1-P(X=0)$

$$=1-\frac{e^{-1}1^0}{1!}$$

$$=0.6321$$

Example 9

If a Poisson distribution is such that $\frac{3}{2}P(X=1)=P(X=3)$, *find (i) $P(X\ge 1)$, (ii) $P(X\le 3)$, and (iii) $P(2\le X\le 5)$.*

Solution

$$\frac{3}{2}P(X=1)=P(X=3)$$

$$\frac{3}{2}\frac{e^{-\lambda}\lambda^1}{1!}=\frac{e^{-\lambda}\lambda^3}{3!}$$

$$\frac{3}{2}=\frac{\lambda^2}{6}$$

$$\lambda^2=9$$

$$\lambda=\pm 3$$

Since $\lambda>0$, $\lambda=3$

$$P(X=x)=\frac{e^{-3}3^x}{x!}, \qquad x=0,1,2,\ldots$$

(i) $P(X\geq 1)=1-P(X<1)$

$$=1-P(X=0)$$

$$=1-\frac{e^{-3}3^0}{0!}$$

$$=0.9502$$

(ii) $P(X\leq 3)=P(X=0)+P(X=1)+P(X=2)+P(X=3)$

$$=\sum_{x=0}^{3}P(X=x)$$

$$=\sum_{x=0}^{3}\frac{e^{-3}3^x}{x!}$$

$$=0.6472$$

(iii) $P(2\leq X\leq 5)=P(X=2)+P(X=3)+P(X=4)+P(X=5)$

$$=\sum_{x=2}^{5}P(X=x)$$

$$=\sum_{x=2}^{5}\frac{e^{-3}3^x}{x!}$$

$$=0.7169$$

Example 10

If X is a Poisson variate such that

$$3\,P(X=4)=\frac{1}{2}P(X=2)+P(X=0)$$

Find (i) the mean of X, and (ii) $P(X\leq 2)$.

Solution

(i) For a Poisson distribution,

$$P(X=x) =\frac{e^{-\lambda}\lambda^x}{x!}, \quad x=0,1,2,\ldots$$

$$3\,P(X=4)=\frac{1}{2}P(X=2)+P(X=0)$$

$$3\frac{e^{-\lambda}\lambda^4}{4!} =\frac{1}{2}\frac{e^{-\lambda}\lambda^2}{2!}+\frac{e^{-\lambda}\lambda^0}{0!}$$

$$\lambda^4-2\lambda^2-8=0$$

$$(\lambda^2-4)(\lambda^2+2)=0$$

$$\lambda=\pm 2 \qquad (\because \lambda \text{ is real})$$

$$\lambda=2 \qquad (\because \lambda>0)$$

Mean $=\lambda=2$

Hence, $P(X=x)=\dfrac{e^{-2}2^x}{x!}, \qquad x=0,1,2,\ldots$

(ii) $P(X\leq 2)=P(X=0)+P(X=1)+P(X=2)$

$$=\sum_{x=0}^{2}P(X=x)$$

$$=\sum_{x=0}^{2}\frac{e^{-2}2^x}{x!}$$

$$=0.6766$$

Example 11

A manufacturer of cotterpins knows that 5% of his products are defective. If he sells cotterpins in boxes of 100 and guarantees that not more than 10 pins will be defective, what is the approximate probability that a box will fail to meet the guaranteed quality?

Solution

Let p be the probability of a pin being defective.

$$p=5\%=0.05, \qquad n=100$$

Since p is very small and n is large, Poisson distribution is used.

$$\lambda=np=100\times 0.05=5$$

Let X be the random variable which denotes the number of defective pins in a box of 100.

Probability of x defective pins in a box of 100

$$P(X=x)=\frac{e^{-\lambda}\lambda^x}{x!}=\frac{e^{-5}5^x}{x!}, \quad x=0,1,2,\ldots$$

Probability that a box will fail to meet the guaranteed quality

$$\begin{aligned}P(X>10)&=1-P(X\le 10)\\&=1-\sum_{x=0}^{10}P(X=x)\\&=1-\sum_{x=0}^{10}\frac{e^{-5}5^x}{x!}\\&=0.0137\end{aligned}$$

Example 12

A car-hire firm has two cars, which it hires out day by day. The number of demands for a car on each day is distributed as a Poisson distribution with a mean of 1.5. Calculate the proportion of days on which (i) neither car is used, and (ii) the proportion of days on which some demand is refused.

Solution

$$\lambda=1.5$$

Let X be the random variable which denotes the number of demands for a car on each day.

Probability of days on which there are x demands for a car

$$P(X=x)=\frac{e^{-\lambda}\lambda^x}{x!}=\frac{e^{-1.5}1.5^x}{x!}, \quad x=0,1,2,\ldots$$

(i) Proportion or probability of days on which neither car is used

$$P(X=0)=\frac{e^{-1.5}1.5^0}{0!}=0.2231$$

(ii) Proportion or probability of days on which some demand is refused

$$\begin{aligned}P(X>2)&=1-P(X\le 2)\\&=1-\sum_{x=0}^{2}P(X=x)\\&=1-\sum_{x=0}^{2}\frac{e^{-1.5}1.5^x}{x!}\\&=0.1912\end{aligned}$$

Example 13

Six coins are tossed 6400 times. Using the Poisson distribution, what is the approximate probability of getting six heads 10 times?

Solution

Let p be the probability of getting one head with one coin.

$$p=\frac{1}{2}$$

Probability of getting 6 heads with 6 coins $=\left(\frac{1}{2}\right)^6=\frac{1}{64}$

$$n=6400$$

$$\lambda=np=6400\left(\frac{1}{64}\right)=100$$

Probability of getting x heads

$$P(X=x)=\frac{e^{-\lambda}\lambda^x}{x!}=\frac{e^{-100}100^x}{x!}, \quad x=0,1,2,\ldots$$

Probability of getting 6 heads 10 times

$$P(X=10)=\frac{e^{-100}100^{10}}{10!}=1.025\times10^{-30}$$

Example 14

If 2% of lightbulbs are defective, find the probability that (i) at least one is defective, and (ii) exactly 7 are defective. Also, find $P(1<X<8)$ *in a sample of* 100.

Solution

Let p be the probability of defective bulb.

$$p=2\%=0.02$$

$$n=100$$

Since p is very small and n is large, Poisson distribution is used.

$$\lambda=np=100(0.02)=2$$

Let X be the random variable which denotes the number of defective bulbs in a sample of 100.

Probability of x defective bulb in a sample of 100

$$P(X=x)=\frac{e^{-\lambda}\lambda^x}{x!}=\frac{e^{-2}2^x}{x!}, \quad x=0,1,2,\ldots$$

(i) Probability that at least one bulb is defective

$$\begin{aligned} P(X \geq 1) &= 1 - P(X=0) \\ &= 1 - \frac{e^{-2}2^0}{0!} \\ &= 0.8647 \end{aligned}$$

(ii) Probability that exactly 7 bulbs are defective

$$P(X=7) = \frac{e^{-2}2^7}{7!} = 0.0034$$

(iii) $P(1 < X < 8) = P(X=2) + P(X=3) + P(X=4) + P(X=5) + P(X=6) + P(X=7)$

$$\begin{aligned} &= \sum_{x=2}^{7} P(X=x) \\ &= \sum_{x=2}^{7} \frac{e^{-2}2^x}{x!} \\ &= 0.5929 \end{aligned}$$

Example 15

An insurance company insured 4000 people against loss of both eyes in a car accident. Based on previous data, the rates were computed on the assumption that on the average, 10 persons in 100000 will have car accidents each year that result in this type of injury. What is the probability that more than 3 of the insured will collect on their policy in a given year?

Solution

Let p be the probability of loss of both eyes in a car accident.

$$p = \frac{10}{100000} = 0.0001$$

$$n = 4000$$

Since p is very small and n is large, Poisson distribution is used.

$$\lambda = np = 4000(0.0001) = 0.4$$

Let X be the random variable which denotes the number of car accidents in a group of 4000 people.

Probability of x car accidents in a group of 4000 people

$$P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-0.4}0.4^x}{x!}, \quad x = 0, 1, 2, \ldots$$

Probability that more than 3 of the insured will collect on their policy, i.e., probability of more than 3 car accidents in a group of 4000 people

$$\begin{aligned} P(X > 3) &= 1 - P(X \leq 3) \\ &= 1 - [P(X=0) + (X=1) + P(X=2) + P(X=3)] \\ &= 1 - \sum_{x=0}^{3} P(X=x) \\ &= 1 - \sum_{x=0}^{3} \frac{e^{-0.4}0.4^x}{x!} \\ &= 0.00077 \end{aligned}$$

Example 16

If two cards are drawn from a pack of 52 cards which are diamonds, using Poisson distribution, find the probability of getting two diamonds at least 3 times in 51 consecutive trials of two cards drawing each time.

Solution

Let p be the probability of getting two diamonds from a pack of 52 cards.

$$p = \frac{^{13}C_2}{^{52}C_2} = \frac{3}{51}, \quad n = 51$$

Since p is very small and n is large, Poisson distribution is used.

$$\lambda = np = 51\left(\frac{3}{51}\right) = 3$$

Let X be the random variable which denotes the drawing of two diamond cards.

Probability of x trials of drawing two diamond cards in 51 trials

$$P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-3}3^x}{x!}, \qquad x = 0, 1, 2, \ldots$$

Probability of getting two diamond cards at least 3 times in 51 trials

$$\begin{aligned} P(X \geq 3) &= 1 - P(X < 3) \\ &= 1 - [P(X=0) + P(X=1) + P(X=2)] \\ &= 1 - \sum_{x=0}^{2} \frac{e^{-3}3^x}{x!} \\ &= 0.5768 \end{aligned}$$

Example 17

Suppose a book of 585 pages contains 43 typographical errors. If these errors are randomly distributed throughout the book, what is the probability that 10 pages, selected at random, will be free from errors?

Solution

Let p be the probability of errors in a page.

$$p = \frac{43}{585} = 0.0735, \quad n = 10$$

Since p is very small and n is large, Poisson distribution is used.

$$\lambda = np = 10(0.0735) = 0.735$$

Let X be the random variable which denotes the errors in the pages.

Probability of x errors in a page in a book of 585 pages

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-0.735}0.735^x}{x!}, \quad x = 0, 1, 2, \ldots$$

Probability that a random sample of 10 pages will contain no error.

$$P(X = 0) = \frac{e^{-0.735}0.735^0}{0!} = 0.4795$$

Example 18

A hospital switchboard receives an average of 4 emergency calls in a 10-minute interval. What is the probability that (i) there are at most 2 emergency calls? (ii) there are exactly 3 emergency calls in an interval of 10 minutes?

Solution

Let p be the probability of receiving emergency calls per minute.

$$p = \frac{4}{10} = 0.4, \quad n = 10$$

$$\lambda = np = 10(0.4) = 4$$

Let X be the random variable which denotes the number of emergency calls per minute.

Probability of x emergency calls per minute

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-4}4^x}{x!}, \quad x = 0, 1, 2, \ldots$$

Probability that there are at most 2 emergency calls

$$\begin{aligned} P(X \le 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= \sum_{x=0}^{2} P(X = x) \\ &= \sum_{x=0}^{2} \frac{e^{-4}4^x}{x!} \\ &= 0.238 \end{aligned}$$

Probability that there are exactly 3 emergency calls

$$P(X = 3) = \frac{e^{-4}4^3}{3!} = 0.1954$$

Example 19

A manufacturer, who produces medicine bottles, finds that 0.1% of the bottles are defective. The bottles are packed in boxes containing 500 bottles. A drug manufacturer buys 100 boxes from the producer of bottles. Using Poisson distribution, find how many boxes will contain (i) no defective bottles and (ii) at least 2 defective bottles.

Solution

Let p be the probability of deflective bottles.

$$p = 0.1\% = 0.001$$

$$n = 500$$

$$\lambda = np = 500(0.001) = 0.5$$

Let X be the random variable which denotes the number of defective bottles in a box of 500.

Probability of x defective bottles in a box of 500

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-0.5}0.5^x}{x!}, \quad x = 0, 1, 2, \ldots$$

(i) Probability of no defective bottles in a box

$$P(X = 0) = \frac{e^{-0.5}0.5^0}{0!} = 0.6065$$

Number of boxes containing no defective bottles

$$f(x) = N\,P(x = 0) = 100(0.6065) \approx 61$$

(ii) Probability of at least 2 defective bottles

$$\begin{aligned} P(X \ge 2) &= 1 - P(X < 2) \\ &= 1 - [P(X = 0) + P(X = 1)] \\ &= 1\sum_{x=0}^{1} P(X = x) \\ &= 1 - \sum_{x=0}^{1} \frac{e^{-0.5}0.5^x}{x!} \\ &= 0.0902 \end{aligned}$$

Number of boxes containing at least 2 defective bottles

$$f(x) = N\,P(X \ge 2) = 100\,(0.0902) \approx 9$$

Example 20

In a certain factory turning out blades, there is a small chance of $\frac{1}{500}$ *for any blade to be defective. The blades are supplied in packets of* 10. *Use the Poisson distribution to calculate the approximate number of packets containing no defective, one defective, and two defective blades in a consignment of* 10000 *packets.*

Solution

Let p be the probability of defective blades in a packet.

$$p=\frac{1}{500}, \quad n=10, \quad N=10000$$

$$\lambda=np=10\left(\frac{1}{500}\right)=0.02$$

Let X be the random variable which denotes the number of defective blades in a packet.

Probability of x defective blades in a packet

$$P(X=x)=\frac{e^{-\lambda}\,\lambda^x}{x!}=\frac{e^{-0.02}\,0.02^x}{x!}, \quad x=0,1,2,\ldots$$

(i) Probability of no defective blades in a packet

$$P(X=0)=\frac{e^{-0.02}\,0.02^0}{0!}=0.9802$$

Number of packets with no defective blades

$$f(x)=N\,P(X=0)=10000(0.9802)=9802$$

(ii) Probability of one defective blade in a packet

$$P(X=1)=\frac{e^{-0.02}\,0.02^1}{1!}=0.0196$$

Number of packets with one defective blade

$$f(x)=N\,P(X=1)=10000\,(0.0196)=196$$

(iii) Probability of two defective blades in a packet

$$P(X=2)=\frac{e^{-0.02}\,0.02^2}{2!}=1.96\times 10^{-4}$$

Number of packets with 2 defective blades

$$f(x)=N\,P(X=2)=10000\,(1.96\times 10^{-4})=1.96\approx 2$$

Example 21

The number of accidents in a year attributed to taxi drivers in a city follows Poisson distribution with a mean of 3. *Out of* 1000 *taxi drivers, find approximately the number of drivers with (i) no accidents in a year, and (ii) more than* 3 *accidents in a year.*

Solution

For a Poisson distribution,

$$\lambda=3, N=1000$$

Probably of x accidents in year

$$P(X=x)=\frac{e^{-\lambda}\,\lambda^x}{x!}=\frac{e^{-3}\,3^x}{x!}, \quad x=0,1,2,\ldots$$

(i) Probability of no accidents in a year

$$P(X=0)=\frac{e^{-3}\,3^0}{0!}=0.0498$$

Number of drivers with no accidents

$$f(x)=N\,P(X=0)=1000(0.0498)=49.8\approx 50$$

(ii) Probability of more than 3 accidents in a year

$$\begin{aligned}P(X>3)&=1-P(X\le 3)\\&=1-[P(X=0)+P(X=1)+P(X=2)+P(X=3)]\\&=1-\sum_{x=0}^{3}P(X=x)\\&=1-\sum_{x=0}^{3}\frac{e^{-3}\,3^x}{x!}\\&=0.3528\end{aligned}$$

Number of drivers with more than 3 accidents

$$f(x)=N\,P(X>3)=1000\,(0.3528)=3528\approx 353$$

Example 22

Fit a Poisson distribution to the following data:

Number of deaths (x)	0	1	2	3	4
Frequency (f)	122	60	15	2	1

Solution

$$\begin{aligned}\text{Mean}&=\frac{\sum fx}{\sum f}\\&=\frac{122(0)+60(1)+15(2)+2(3)+1(4)}{122+60+15+2+1}\\&=\frac{100}{200}\\&=0.5\end{aligned}$$

For a Poisson distribution,

$$\lambda = 0.5$$

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-0.5}\,0.5^x}{x!}, \quad x = 0, 1, 2, 3, 4$$

$$N = \sum f = 100$$

Theoretical or expected frequency $f(x) = N\,P(X = x)$

$$f(x) = \frac{200\,e^{-0.5}\,0.5^x}{x!}$$

$$f(0) = \frac{200\,e^{-0.5}\,0.5^0}{0!} = 121.31 \approx 121$$

$$f(1) = \frac{200\,e^{-0.5}\,0.5^1}{1!} = 60.65 \approx 61$$

$$f(2) = \frac{200\,e^{-0.5}\,0.5^2}{2!} = 15.16 \approx 15$$

$$f(3) = \frac{200\,e^{-0.5}\,0.5^3}{3!} = 2.53 \approx 3$$

$$f(4) = \frac{200\,e^{-0.5}\,0.5^4}{4!} = 0.32 \approx 0$$

Poisson Distribution

Number of deaths (x)	0	1	2	3	4
Expected Poisson frequency $f(x)$	121	61	15	3	0

Example 23

Assuming that the typing mistakes per page committed by a typist follows a Poisson distribution, find the expected frequencies for the following distribution of typing mistakes:

Number of mistakes per page	0	1	2	3	4	5
Number of pages	40	30	20	15	10	5

Solution

$$\text{Mean} = \frac{\sum fx}{\sum f}$$

$$= \frac{40(0) + 30(1) + 20(2) + 15(3) + 10(4) + 5(5)}{40 + 30 + 20 + 15 + 10 + 5}$$

$$= \frac{180}{120}$$

$$= 1.5$$

For a Poisson distribution,

$$\lambda = 1.5$$

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-1.5}\,1.5^x}{x!}, \quad x = 0, 1, 2, 3, 4, 5$$

$$N = \sum f = 120$$

Expected frequency $f(x) = N\,P(X = x)$

$$f(x) = \frac{120\,e^{-1.5}\,1.5^x}{x!}$$

$$f(0) = \frac{120\,e^{-1.5}\,1.5^0}{0!} = 26.78 \approx 27$$

$$f(1) = \frac{120\,e^{-1.5}\,1.5^1}{1!} = 40.16 \approx 40$$

$$f(2) = \frac{120\,e^{-1.5}\,1.5^2}{2!} = 30.12 \approx 30$$

$$f(3) = \frac{120\,e^{-1.5}\,1.5^3}{3!} = 15.06 \approx 15$$

$$f(4) = \frac{120\,e^{-1.5}\,1.5^4}{4!} = 5.65 \approx 6$$

$$f(5) = \frac{120\,e^{-1.5}\,1.5^5}{5!} = 1.69 \approx 2$$

EXERCISE 5.2

1. The mean and variance of a probability distribution is 2. Write down the distribution.

   $\left[\textbf{Ans.:}\ P(X = x) = \frac{e^{-2}2^x}{x!},\ x = 0, 1, 2, \ldots\right]$

2. In a Poisson distribution, the probability $P(X = 0)$ is 20 per cent. Find the mean of the distribution.

   [**Ans.:** 2.9957]

3. If X is a Poisson variate and $P(X = 0) = 6\,P(X = 3)$, find $P(X = 2)$.

   [**Ans.:** 0.1839]

4. The standard deviation of a Poisson distribution is 3. Find the probability of getting 3 successes.

[**Ans.:** 0.0149]

5. The probability that a Poisson variable X takes a positive value is $1 - e^{-1.5}$. Find the variance and the probability that X lies between -1.5 and 1.5.

[**Ans.:** 1.5, 0.5578]

6. If 2 per cent bulbs are known to be defective bulbs, find the probability that in a lot of 300 bulbs, there will be 2 or 3 defective bulbs using Poisson distribution.

[**Ans.:** 0.1338]

7. In a certain manufacturing process, 5% of the tools produced turn out to be defective. Find the probability that in a sample of 40 tools, at most 2 will be defective.

[**Ans.:** 0.675]

8. If the probability that an individual suffers a bad reaction from a particular injection is 0.001, determine the probability that out of 2000 individuals (i) exactly three, and (ii) more than two individuals suffer a bad reaction.

[**Ans.:** (i) 0.1804 (ii) 0.3233]

9. It is known from past experience that in a certain plant, there are on the average 4 industrial accidents per year. Find the probability that in a given year, there will be less than 4 accidents. Assume Poisson distribution.

[**Ans.:** 0.43]

10. Find the probability that at most 5 defective fuses will be found in a box of 200 fuses, if experience shows that 2% of such fuses are defective.

[**Ans.:** 0.7851]

11. Assume that the probability of an individual coal minor being killed in a mine accident during a year is $\frac{1}{2400}$. Use appropriate statistical distribution to calculate the probability that in a mine employing 200 miners, there will be at least one fatal accident every year.

[**Ans.:** 0.07]

12. Between the hours of 2 and 4 p.m., the average number of phone calls per minute coming into the switchboard of a company is 2.5. Find the probability that during a particular minute, there will be (i) no phone call at all, (ii) 4 or less calls, and (iii) more than 6 calls.

[**Ans.:** (i) 0.0821 (ii) 0.8909 (iii) 0.0145]

13. Suppose that a local appliances shop has found from experience that the demand for tubelights is roughly distributed as Poisson with a mean of 4 tubelights per week. If the shop keeps 6 tubelights during a particular week, what is the probability that the demand will exceed the supply during that week?

[**Ans.:** 0.1106]

14. The distribution of the number of road accidents per day in a city is Poisson with a mean of 4. Find the number of days out of 100 days when there will be (i) no accident, (i) at least 2 accidents, and (iii) at most 3 accidents.

[**Ans.:** (i) 2 (ii) 91 (iii) 44]

15. A manufacturer of electric bulbs sends out 500 lots each consisting of 100 bulbs. If 5% bulbs are defective, in how many lot can we expect (i) 97 or more good bulbs? (ii) less than 96 good bulbs?

[**Ans.:** (i) 62 (ii) 132]

16. A firm produces articles, 0.1 per cent of which are defective. It packs them in cases containing 500 articles. If a wholesaler purchases 100 such cases, how many cases can be expectecd (i) to be free from defects? (ii) to have one defective article?

[**Ans.:** (i) 16 (ii) 30]

17. In a certain factory producing certain articles, the probability that an article is defective is $\frac{1}{500}$. The articles are supplied in packets of 20. Find approximately the number of packets containing no defective, one defective, two defectives in a consignment of 20000 packets.

[**Ans.:** 19200, 768, 15]

18. In a certain factory manufacturing razor blades, there is a small chance, $\frac{1}{50}$ for any blade to be defective. The blades are placed in packets, each containing 10 blades. Using the Poisson distribution, calculate the approximate number of packets containing not more than 2 defective blades in a consignment of 10000 packets.

[**Ans.:** 9988]

19. It is known that 0.5% of ballpen refills produced by a factory are defective. These refills are dispatched in packaging of equal numbers. Using a Poisson distribution, determine the number of refills in a packing to be sure that at least 95% of them contain no defective refills.

[Ans.: 10]

20. A manufacturer finds that the average demand per day for the mechanics to repair his new product is 1.5 over a period of one year and the demand per day is distributed as a Poisson variate. He employs two mechanics. On how many days in one year (i) would both mechanics be free? (ii) some demand is refused?

[Ans.: (i) 81.4 days (ii) 69.8 days]

21. Fit a Poisson distribution to the following data:

X	0	1	2	3	4
f	211	90	19	5	0

[Ans.: $\lambda = 0.44$, Frequencies : 209, 92, 20, 3, 1]

22. Fit a Poisson distribution to the following data:

No. of defects per piece	0	1	2	3	4
No. of pieces	43	40	25	10	2

[Ans.: Frequencies: 42, 44, 24, 8, 2]

23. Fit a Poisson distribution to the following data:

X	0	1	2	3	4	5
f	142	156	69	27	5	1

[Ans.: Frequencies: 147, 147, 74, 24, 6, 2]

24. Fit a Poisson distribution to the following data:

X	0	1	2	3	4	5	6	7	8
f	56	156	132	92	37	22	4	0	1

[Ans.: Frequency : 70, 137, 135, 89, 44, 17, 6, 2, 0]

5.4 NORMAL DISTRIBUTION

A continuous random variable X is said to follow normal distribution with mean μ and variance σ^2, if its probability function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$-\infty < X < \infty, -\infty < \mu < \infty, \sigma > 0$$

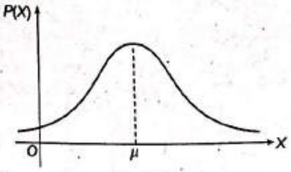


Fig. 5.1

where μ and σ are called parameters of the normal distribution. The curve representing the normal distribution is called the normal curve (Fig. 5.1).

5.4.1 Properties of the Normal Distribution

A normal probability curve, or normal curve, has the following properties:

(i) It is a bell-shaped symmetrical curve about the ordinate $X = \mu$. The ordinate is maximum at $X = \mu$.

(ii) It is a unimodal curve and its tails extend infinitely in both the directions, i.e., the curve is asymptotic to X-axis in both the directions.

(iii) All the three measures of central tendency coincide, i.e.,

$$\text{mean} = \text{median} = \text{mode}$$

(iv) The total area under the curve gives the total probability of the random variable X taking values between $-\infty$ to ∞. Mathematically,

$$P(-\infty < X < \infty) = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \mathrm{d}x = 1$$

(v) The ordinate at $X = \mu$ divides the area under the normal curve into two equal parts, i.e.,

$$\int_{-\infty}^{\mu} f(x)\mathrm{d}x = \int_{\mu}^{\infty} f(x)\mathrm{d}x = \frac{1}{2}$$

(vi) The value of $f(x)$ is always nonnegative for all values of X, i.e., the whole curve lies above the X-axis.

(vii) The points of inflexion (the point at which curvature changes) of the curve are at $X = \mu + \sigma$ and the curve changes from concave to convex at $X = \mu + \sigma$ to $X = \mu - \sigma$.

(viii) The area under the normal curve (Fig. 5.2) is distributed as follows:

(a) The area between the ordinates at $\mu - \sigma$ and $\mu + \sigma$ is 68.27%

(b) The area between the ordinates at $\mu - 2\sigma$ and $\mu + 2\sigma$ is 95.45%

(c) The area between the ordinates at $\mu - 3\sigma$ and $\mu + 3\sigma$ is 99.74%

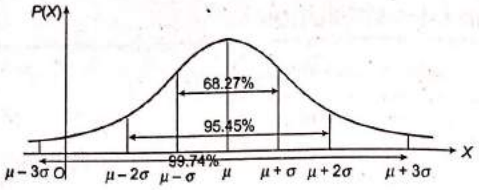


Fig. 5.2

5.4.2 Constants of the Normal Distribution

1. Mean of the Normal Distribution

$$E(X) = \int_{-\infty}^{\infty} x\, f(x)\,\mathrm{d}x$$

$$= \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \mathrm{d}x$$

Putting $\frac{x-\mu}{\sigma} = t,\ \mathrm{d}x = \sigma\,\mathrm{d}t$

$$E(X) = \int_{-\infty}^{\infty} (\mu + \sigma t) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}\,\mathrm{d}t$$

$$= \mu \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}\,\mathrm{d}t + \int_{-\infty}^{\infty} \sigma \frac{t}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}\,\mathrm{d}t$$

Putting $t^2 = u$ in the second integral,

$2t\,\mathrm{d}t = \mathrm{d}u$

When $t \to \infty, \quad u \to \infty$

When $t \to -\infty, \quad u \to \infty$

$$E(X) = \mu \frac{1}{\sqrt{2\pi}} \cdot \sqrt{2\pi} + \int_{\infty}^{\infty} \sigma \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u} \frac{\mathrm{d}u}{2} \qquad \left[\because \int_{-\infty}^{\infty} e^{-\frac{1}{2}t^2}\,\mathrm{d}t = \sqrt{2\pi}\right]$$

$$= \mu + 0 \qquad [\because \text{ the limits of integration are same}]$$

$$= \mu$$

2. Variance of the Normal Distribution

$$\mathrm{Var}(X) = E(X - \mu)^2$$

$$= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,\mathrm{d}x$$

$$= \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \mathrm{d}x$$

Putting $\frac{x-\mu}{\sigma} = t,\ \mathrm{d}x = \sigma\,\mathrm{d}t$

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} \sigma^2 t^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}\,\mathrm{d}t$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^2 e^{-\frac{1}{2}t^2}\,\mathrm{d}t$$

$$= \frac{2\sigma^2}{\sqrt{2\pi}} \int_{0}^{\infty} t^2 e^{-\frac{1}{2}t^2}\,\mathrm{d}t \qquad [\because \text{ integral is an even function}]$$

Putting $\frac{t^2}{2} = u$,

$t = \sqrt{2u}$

$\mathrm{d}t = \sqrt{2} \frac{1}{2\sqrt{u}}\,\mathrm{d}u = \frac{1}{\sqrt{2u}}\,\mathrm{d}u$

When $t = 0, \quad u = 0$

When $t = \infty, \quad u = \infty$

$$\mathrm{Var}(X) = \frac{2\sigma^2}{\sqrt{2\pi}} \int_0^{\infty} 2u e^{-u} \frac{1}{\sqrt{2u}}\,\mathrm{d}u$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} \int_0^{\infty} e^{-u} u^{\frac{1}{2}}\,\mathrm{d}u$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} \left\lceil \frac{3}{2} \right. \qquad \left[\because \int_0^{\infty} e^{-x} x^{n-1}\,\mathrm{d}x = \left\lceil n \right.\right]$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} \frac{1}{2} \left\lceil \frac{1}{2} \right.$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} \frac{1}{2} \sqrt{\pi}$$

$$= \sigma^2$$

3. Standard Deviation of the Normal Distribution

$$\mathrm{SD} = \sigma$$

4. Mode of the Normal Distribution

Mode is the value of x for which $f(x)$ is maximum. Mode is given by

$$f'(x)=0 \text{ and } f''(x)<0$$

For normal distribution,

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Differentiating w.r.t. x,

$$f'(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\left[-\left(\frac{x-\mu}{\sigma^2}\right)\right]$$

$$=-\frac{x-\mu}{\sigma^2}f(x)$$

When $f'(x)=0, \quad x-\mu=0$

$$x=\mu$$

$$f''(x)=-\frac{1}{\sigma^2}[(x-\mu)f'(x)+f(x)]$$

$$=-\frac{1}{\sigma^2}\left[(x-\mu)\left\{-\frac{(x-\mu)}{\sigma^2}f(x)\right\}+f(x)\right]$$

$$=-\frac{1}{\sigma^2}f(x)\left[1-\frac{(x-\mu)^2}{\sigma^2}\right]$$

At $x=\mu$,

$$f''(x)=\frac{f(x)}{\sigma^2}=-\frac{1}{\sigma^3\sqrt{2\pi}}<0$$

Hence, $x=\mu$ is the mode of the normal distribution.

5. Median of the Normal Distribution

If M is median of the normal distribution,

$$\int_{-\infty}^{M} f(x)\,dx=\frac{1}{2}$$

$$\int_{-\infty}^{M}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx=\frac{1}{2}$$

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\mu}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx+\frac{1}{\sigma\sqrt{2\pi}}\int_{\mu}^{M}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx=\frac{1}{2} \quad \text{...(5.3)}$$

Putting $\frac{x-\mu}{\sigma}=t$ in the first integral,

$$dx=\sigma\,dt$$

When $x=-\infty, \quad t=-\infty$

When $x=\mu, \quad t=0$

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\mu}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{0}e^{-\frac{1}{2}t^2}\sigma\,dt$$

$$=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0}e^{-\frac{1}{2}t^2}dt$$

$$=\frac{1}{\sqrt{2\pi}}\int_{0}^{\infty}e^{-\frac{1}{2}t^2}dt \quad \text{[By symmetry]}$$

$$=\frac{1}{\sqrt{2\pi}}\sqrt{\frac{\pi}{2}}$$

$$=\frac{1}{2} \quad \text{...(5.4)}$$

From Eqs (5.3) and (5.4),

$$\frac{1}{2}+\frac{1}{\sigma\sqrt{2\pi}}\int_{\mu}^{M}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx=\frac{1}{2}$$

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{\mu}^{M}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx=0$$

$$\int_{\mu}^{M}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx=0$$

$$\mu=M\left[\because \text{ if } \int_{a}^{b}f(x)\,dx=0 \text{ then } a=b \text{ where } f(x)>0\right]$$

Hence, mean = median for the normal distribution.

Note For normal distribution,

mean = median = mode = μ

Hence, the normal distribution is symmetrical.

5.4.3 Probability of a Normal Random Variable in an Interval

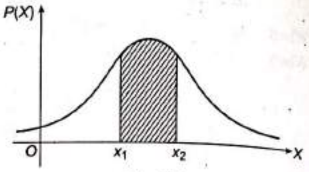


Fig. 5.3

Let X be a normal random variable with mean μ and standard deviation σ. The probability of X lying in the interval (x_1, x_2) (Fig. 5.3) is given by

$$P(x_1 \le X \le x_2) = \int_{x_1}^{x_2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Hence, the probability is equal to the area under the normal curve between the ordinates $X = x_1$ and $X = x_2$ respectively. $P(x_1 < X < x_2)$ can be evaluated easily by converting a normal random variable into another random variable.

Let $\quad Z = \frac{X-\mu}{\sigma}$ be a new random variable.

$$E(Z) = E\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma}[E(X)-\mu] = 0$$

$$\text{Var}(Z) = \text{Var}\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2}\text{Var}(X-\mu) = \frac{1}{\sigma^2}\text{Var}(X) = 1$$

The distribution of Z is also normal. Thus, if X is a normal random variable with mean μ and standard deviation σ then $Z = \frac{X-\mu}{\sigma}$ is a normal random variable with mean 0 and standard deviation 1. Since the parameters of the distribution of Z are fixed, it is a known distribution and is termed *standard normal distribution*. Further, Z is termed as a *standard normal variate*. Thus, the distribution of any normal variate X can always be transformed into the distribution of the standard normal variate Z.

$$P(x_1 \le X \le x_2) = P\left[\left(\frac{x_1-\mu}{\sigma}\right) \le \left(\frac{X-\mu}{\sigma}\right) \le \left(\frac{x_2-\mu}{\sigma}\right)\right]$$
$$= P(z_1 \le Z \le z_2)$$

where $\quad z_1 = \frac{x_1-\mu}{\sigma}$ and $z_2 = \frac{x_2-\mu}{\sigma}$

This probability is equal to the area under the standard normal curve between the ordinates at $Z = z_1$ and $Z = z_2$.

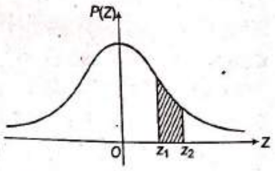


Fig. 5.4

Case I If both z_1 and z_2 are positive (or both negative) (Fig. 5.4),

$$\begin{aligned} P(x_1 \le X \le x_2) &= P(z_1 \le Z \le z_2) \\ &= P(0 \le Z \le z_2) - P(0 \le Z \le z_1) \\ &= (\text{Area under the normal curve from 0 to } z_2) \\ &\quad - (\text{Area under the normal curve from 0 to } z_1) \end{aligned}$$

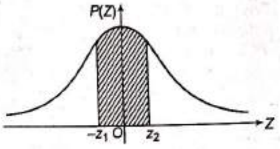


Fig. 5.5

Case II If $z_1 < 0$ and $z_2 > 0$ (Fig. 5.5),

$$\begin{aligned} P(x_1 \le X \le x_2) &= P(-z_1 \le Z \le z_2) \\ &= P(-z_1 \le Z \le 0) + P(0 \le Z \le z_2) \\ &= P(0 \le Z \le z_1) + P(0 \le Z \le z_2) \quad \text{[By symmetry]} \\ &= (\text{Area under the normal curve from 0 to } z_1) \\ &\quad + (\text{Area under the normal curve from 0 to } z_2) \end{aligned}$$

When $X > x_1$, $Z > z_1$, the probability $P(Z > z_1)$ can be found for two cases as follows:

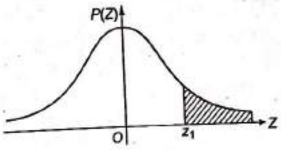


Fig. 5.6

Case I If $z_1 > 0$ (Fig. 5.6),

$$\begin{aligned} P(X > x_1) &= P(Z > z_1) \\ &= 0.5 - P(0 \le Z \le z_1) \\ &= 0.5 - (\text{Area under the curve from 0 to } z_1) \end{aligned}$$

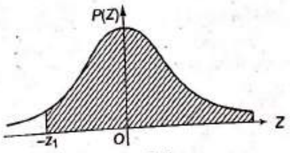


Fig. 5.7

Case II If $z_1 < 0$ (Fig. 5.7),

$$\begin{aligned} P(X > x_1) &= P(Z > -z_1) \\ &= 0.5 + P(-z_1 < Z < 0) \\ &= 0.5 + P(0 < Z < z_1) \quad \text{[By symmetry]} \\ &= 0.5 + (\text{Area under the curve from 0 to } z_1) \end{aligned}$$

When $X < x_1$, $Z < z_1$, the probability $P(Z < z_1)$ can be found for two cases as follows:

Case I If $z_1 > 0$ (Fig. 5.8),

$$\begin{aligned} P(X < x_1) &= P(Z < z_1) \\ &= 1 - P(Z \ge z_1) \\ &= 1 - [0.5 - P(0 < Z < z_1)] \\ &= 0.5 + P(0 < Z < z_1) \\ &= 0.5 + (\text{Area under the curve from 0 to } z_1) \end{aligned}$$

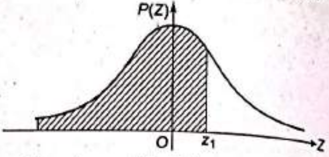


Fig. 5.8

Case II If $z_1 < 0$ (Fig. 5.9),

$$\begin{aligned} P(X < x_1) &= P(Z < -z_1) \\ &= 1 - P(Z \ge -z_1) \\ &= 1 - [0.5 + P(-z_1 \le Z \le 0)] \\ &= 1 - [0.5 + P(0 \le Z \le z_1)] \quad \text{[By symmetry]} \\ &= 0.5 - P(0 \le Z \le z_1) \\ &= 0.5 - (\text{Area under the curve from 0 to } z_1) \end{aligned}$$

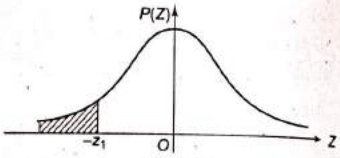


Fig. 5.9

Note

(i) $P(X < x_1) = F(x_1) = \int_{-\infty}^{x_1} f(x)\,dx$

Hence, $P(X < x_1)$ represents the area under the curve from $X = -\infty$ to $X = x_1$.

(ii) If $P(X < x_1) < 0.5$, the point x_1 lies to the left of $X = \mu$ and the corresponding value of standard normal variate will be negative (Fig. 5.10).

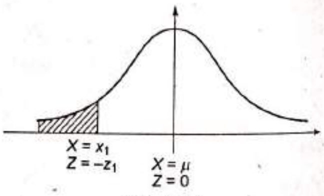


Fig. 5.10

(iii) If $P(X < x_1) > 0.5$, the point x_1 lies to the right of $x = \mu$ and the corresponding value of standard normal variate will be positive (Fig. 5.11).

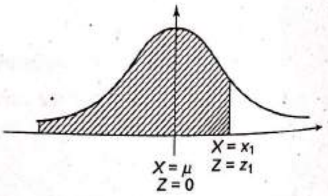
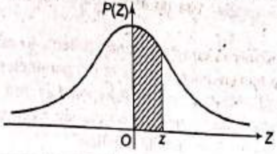


Fig. 5.11

Standard Normal (Z) Table, Area between 0 and z



Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3990	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4115	0.4131	0.4147	0.4162
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

5.4.4 Uses of Normal Distribution

(i) The normal distribution can be used to approximate binomial and Poisson distributions.

(ii) It is used extensively in sampling theory. It helps to estimate parameters from statistics and to find confidence limits of the parameter.

(iii) It is widely used in testing statistical hypothesis and tests of significance in which it is always assumed that the population from which the samples have been drawn should have normal distribution.

(iv) It serves as a guiding instrument in the analysis and interpretation of statistical data.

(v) It can be used for smoothing and graduating a distribution which is not normal simply by contracting a normal curve.

Example 1

What is the probability that a standard normal variate Z will be (i) greater than 1.09*? (ii) less than* −1.65*? (iii) lying between* −1 *and* 1.96*? (iv) lying between* 1.25 *and* 2.75*?*

Solution

(i) $Z > 1.09$ (Fig. 5.12)

$$\begin{aligned} P(Z>1.09) &= 0.5 - P(0 \le Z \le 1.09) \\ &= 0.5 - 0.3621 \\ &= 0.1379 \end{aligned}$$

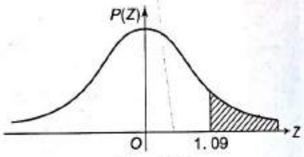


Fig. 5.12

(ii) $Z \le -1.65$ (Fig. 5.13)

$$\begin{aligned} P(Z \le -1.65) &= 1 - P(Z > -1.65) \\ &= 1 - [0.5 + P(-1.65 < Z < 0)] \\ &= 1 - [0.5 + P(0 < Z < 1.65)] \quad \text{[By symmetry]} \\ &= 0.5 - P(0 < Z < 1.65) \\ &= 0.5 - 0.4505 \\ &= 0.0495 \end{aligned}$$

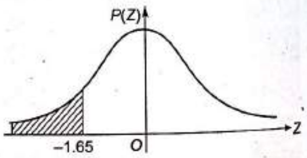


Fig. 5.13

(iii) $-1 < Z < 1.96$ (Fig. 5.14)

$$\begin{aligned} P(-1 < Z < 1.96) &= P(-1 < Z < 0) + P(0 < Z < 1.96) \\ &= P(0 < Z < 1) + P(0 < Z < 1.96) \quad \text{[By symmetry]} \\ &= 0.3413 + 0.4750 \\ &= 0.8163 \end{aligned}$$

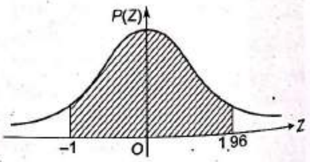


Fig. 5.14

(iv) $1.25 < Z < 2.75$ (Fig. 5.15)

$$\begin{aligned} &P(1.25 < Z < 275) \\ &= P(0 < Z < 2.75) - P(0 < Z < 1.25) \\ &= 0.4970 - 0.3944 \\ &= 0.1026 \end{aligned}$$

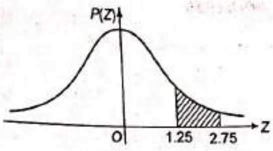


Fig. 5.15

Example 2

If X is a normal variate with a mean of 30 *and an SD of* 5*, find the probabilities that (i)* $26 \le X \le 40$*, and (ii)* $X \ge 45$.

Solution

$\mu = 30, \quad \sigma = 5$

$$Z = \frac{X-\mu}{\sigma}$$

(i) When $X = 26$, $Z = \dfrac{26-30}{5} = -0.8$

When $X = 40$, $Z = \dfrac{40-30}{5} = 2$

$$\begin{aligned} P(26 \le X \le 40) &= P(-0.8 \le Z \le 2) \text{ (Fig. 5.16)} \\ &= P(-0.8 \le Z \le 0) + P(0 \le Z \le 2) \\ &= P(0 \le Z \le 0.8) + P(0 \le Z \le 2) \quad \text{[By symmetry]} \\ &= 0.2881 + 0.4772 \\ &= 0.7653 \end{aligned}$$

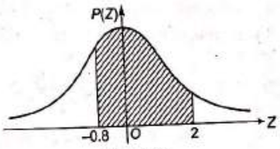


Fig. 5.16

(ii) When $X = 45$, $Z = \dfrac{45-30}{5} = 3$

$$\begin{aligned} P(X \ge 45) &= P(Z \ge 3) \text{ (Fig. 5.17)} \\ &= 0.5 - P(0 < Z < 3) \\ &= 0.5 - 0.4987 \\ &= 0.0013 \end{aligned}$$

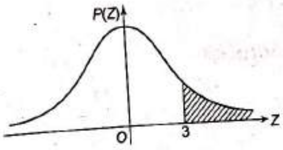


Fig. 5.17

Example 3

X is normally distributed and the mean of X is 12 *and the SD is* 4*. Find out the probability of the following:*

(i) $X \ge 20$ *(ii)* $X \le 20$ *(iii)* $0 \le X \le 12$.

Solution

$$\mu = 12, \quad \sigma = 4$$

$$Z = \frac{X-\mu}{\sigma}$$

(i) When $X = 20, Z = \frac{20-12}{4} = 2$

$$P(X \geq 20) = P(Z \geq 2) \text{ (Fig. 5.18)}$$
$$= 0.5 - P(0 < Z < 2)$$
$$= 0.5 - 0.4772$$
$$= 0.0228$$

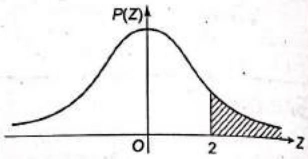


Fig. 5.18

(ii) $P(X \leq 20) = 1 - P(X > 20)$

$$= 1 - 0.0228$$
$$= 0.9772$$

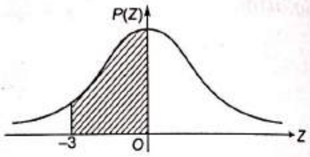


Fig. 5.19

(iii) When $X = 0, Z = \frac{0-12}{4} = -3$

When $X = 12, Z = \frac{12-12}{4} = 0$

$$P(0 \leq X \leq 12) = P(-3 \leq Z \leq 0) \text{ (Fig. 5.19)}$$
$$= P(0 \leq Z \leq 3) \qquad \text{[By symmetry]}$$
$$= 0.4987$$

Example 4

If X is normally distributed with a mean of 2 and an SD of 0.1, find $P(|X-2|) \geq 0.01)$?

Solution:

$$\mu = 2, \quad \sigma = 0.1$$

$$Z = \frac{X-\mu}{\sigma}$$

When $X = 1.99, Z = \frac{1.99-2}{0.1} = -0.1$

When $X = 2.01, Z = \frac{2.01-2}{0.1} = 0.1$

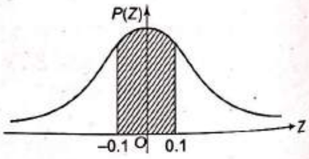


Fig. 5.20

$$P(|X-2| \leq 0.01) = P(1.99 \leq X \leq 2.01) \text{ (Fig. 5.20)}$$
$$= P(-0.1 \leq Z \leq 0.1)$$
$$= P(-0.1 \leq Z \leq 0) + P(0 \leq Z \leq 0.1)$$
$$= P(0 \leq Z \leq 0.1) + P(0 \leq Z \leq 0.1) \qquad \text{[By symmetry]}$$
$$= 2P(0 < Z \leq 0.1)$$
$$= 2(0.0398)$$
$$= 0.0796$$

$$P(|X-2| \geq 0.01) = 1 - P(|X-2| < 0.01)$$
$$= 1 - 0.0796$$
$$= 0.9204$$

Example 5

If X is a normal variate with a mean of 120 and a standard deviation of 10, find c such that (i) P(X > c) = 0.02, and (ii) P(X< c) = 0.05.

Solution

For normal variate X,

$$\mu = 120, \quad \sigma = 10$$

$$Z = \frac{X-\mu}{\sigma}$$

(i) $P(X > c) = 0.02$

$$P(X < c) = 1 - P(X \geq c)$$
$$= 1 - 0.02$$
$$= 0.98$$

Since $P(X < c) > 0.5$, the corresponding value of Z will be positive.

$$P(X > c) = P(Z > z_1) \text{ (Fig. 5.21)}$$
$$0.02 = 0.5 - P(0 \leq Z \leq z_1)$$
$$P(0 \leq Z \leq z_1) = 0.48$$
$$\therefore \; z_1 = 2.05 \quad \text{[From normal table]}$$

$$Z = \frac{c-120}{10} = z_1 = 2.05$$
$$c = 2.05(10) + 120 = 140.05$$

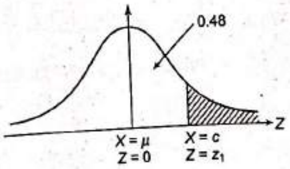


Fig. 5.21

(ii) Since $P(X<c)<0.5$, the corresponding value of Z will be negative.

$$P(X < c) = P(Z < -z_1) \text{ (Fig. 5.22)}$$
$$0.05 = 1 - P(Z \geq -z_1)$$
$$0.05 = 1 - [0.5 + P(-z_1 \leq Z \leq 0)]$$

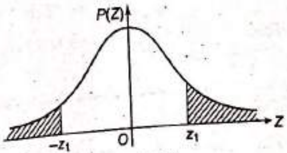


Fig. 5.22

$$0.05 = 1 - [0.5 + P(0 \le Z \le z_1)] \quad \text{[By symmetry]}$$
$$0.05 = 0.5 - P(0 \le Z \le z_1)$$
$$P(0 \le Z \le z_1) = 0.5 - 0.05 = 0.45$$
$$\therefore \quad z_1 = -1.64 \quad \text{[From normal table]}$$
$$Z = \frac{c-120}{10} = z_1 = -1.64$$
$$c = 10(-1.64) + 120 = 103.6$$

Example 6

A manufacturer knows from his experience that the resistances of resistors he produces is normal with $\mu = 100$ ohms and SD = $\sigma = 2$ ohms. What percentage of resistors will have resistances between 98 ohms and 102 ohms?

Solution

Let X be the random variable which denotes the resistances of the resistors.

$$\mu = 100, \quad \sigma = 2$$
$$Z = \frac{X-\mu}{\sigma}$$

When $X = 98$, $Z = \frac{98-100}{2} = -1$

When $X = 102$, $Z = \frac{102-100}{2} = 1$

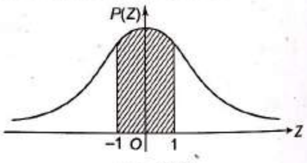


Fig. 5.23

$$\begin{aligned} P(98 \le X \le 102) &= P(-1 \le Z \le 1) \quad \text{(Fig. 5.23)} \\ &= P(-1 \le Z \le 0) + P(0 \le Z \le 1) \\ &= P(0 \le Z \le 1) + P(0 \le Z \le 1) \quad \text{[By symmetry]} \\ &= 2P(0 \le Z \le 1) \\ &= 2(0.3413) \\ &= 0.6826 \end{aligned}$$

Hence, the percentage of resistors have resistances between 98 ohms and 102 ohms = 68.26%.

Example 7

The average seasonal rainfall in a place is 16 inches with an SD of 4 inches. What is the probability that the rainfall in that place will be between 20 and 24 inches in a year?

Solution

Let X be the random variable which denotes the seasonal rainfall in a year.

$$\mu = 16, \quad \sigma = 4$$
$$Z = \frac{X-\mu}{\sigma}$$

When $X = 20$, $Z = \frac{20-16}{4} = 1$

When $X = 24$, $Z = \frac{24-16}{4} = 2$

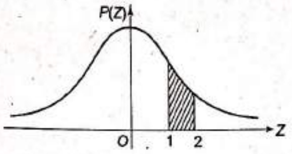


Fig. 5.24

$$\begin{aligned} P(20 < X < 24) &= P(1 < Z < 2) \quad \text{(Fig. 5.24)} \\ &= P(0 < Z < 2) - P(0 < Z < 1) \\ &= 0.4772 - 0.3413 \\ &= 0.1359 \end{aligned}$$

Example 8

The lifetime of a certain kind of batteries has a mean life of 400 hours and the standard deviation as 45 hours. Assuming the distribution of lifetime to be normal, find (i) the percentage of batteries with a lifetime of at least 470 hours, (ii) the proportion of batteries with a lifetime between 385 and 415 hours, and (iii) the minimum life of the best 5% of batteries.

Solution

Let X be the random variable which denotes the lifetime of a certain kind of batteries.

$$\mu = 400, \quad \sigma = 45$$
$$Z = \frac{X-\mu}{\sigma}$$

(i) When $X = 470$,

$$Z = \frac{470-400}{45} = 1.56$$

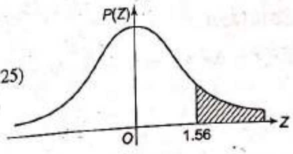


Fig. 5.25

$$\begin{aligned} P(X \ge 470) &= P(Z \ge 1.56) \quad \text{(Fig. 5.25)} \\ &= 0.5 - P(0 < Z < 1.56) \\ &= 0.5 - 0.4406 \\ &= 0.0594 \end{aligned}$$

Hence, the percentage of batteries with a lifetime of at least 470 hours = 5.94%.

(ii) When $X = 385$,

$$Z = \frac{385-400}{45} = -0.33$$

When $X = 415$,

$$Z = \frac{415-400}{45} = 0.33$$

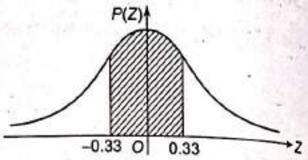


Fig. 5.26

$$\begin{aligned} P(385 < X < 415) &= P(-0.33 < Z < 0.33) \quad \text{(Fig. 5.26)} \\ &= P(-0.33 < Z < 0) + P(0 < Z < 0.33) \\ &= P(0 < Z < 0.33) + P(0 < Z < 0.33) \qquad \text{[By symmetry]} \\ &= 2\,P(0 < Z < 0.33) \\ &= 2(0.1293) \\ &= 0.2586 \end{aligned}$$

Hence, the proportion of batteries with a lifetime between 385 and 415 hours = 25.86%.

(iii) $P(X > x_1) = 0.05$ (Fig. 5.27)

$P(X > x_1) = P(Z > z_1)$

$0.05 = 0.5 - P(0 \le Z \le z_1)$

$P(0 \le Z \le z_1) = 0.5 - 0.05 = 0.45$

$\therefore\ z_1 = 1.65$ [From normal table]

$$Z = \frac{x_1 - 400}{45} = z_1 = 1.65$$

$\therefore\ x_1 = 1.65(45) + 400 = 474.25$ hours

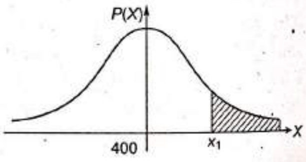


Fig. 5.27

Example 9

If the weights of 300 students are normally distributed with a mean of 68 kg and a standard deviation of 3 kg, how many students have weights (i) greater than 72 kg? (ii) less than or equal to 64 kg? (iii) between 65 kg and 71 kg inclusive?

Solution

Let X be the random variable which denotes the weight of a student.

$\mu = 68, \quad \sigma = 3, \quad N = 300$

$$Z = \frac{X - \mu}{\sigma}$$

(i) When $X = 72$, $Z = \dfrac{72-68}{3} = 1.33$

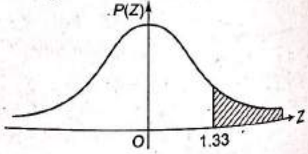


Fig. 5.28

$$\begin{aligned} P(X > 72) &= P(Z > 1.33) \quad \text{(Fig. 5.28)} \\ &= 0.5 - P(0 \le Z \le 1.33) \\ &= 0.5 - 0.4082 \\ &= 0.0918 \end{aligned}$$

$$\begin{aligned} \text{Number of students with weights more than 72 kg} &= N\,P(X > 72) \\ &= 300(0.0918) \\ &= 27.54 \\ &\approx 28 \end{aligned}$$

(ii) When $X = 64$, $Z = \dfrac{64-68}{3} = -1.33$

$$\begin{aligned} P(X \le 64) &= P(Z \le -1.33) \quad \text{(Fig. 5.29)} \\ &= P(Z \ge 1.33) \qquad \text{[By symmetry]} \\ &= 0.5 - P(0 < Z < 1.33) \\ &= 0.5 - 0.4082 \\ &= 0.0918 \end{aligned}$$

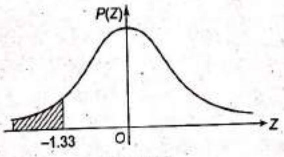


Fig. 5.29

Number of students with weights less than or equal to 64 kg

$$\begin{aligned} &= N\,P(X \le 64) \\ &= 300(0.0918) \\ &= 27.54 \\ &\approx 28 \end{aligned}$$

(iii) When $X = 65$, $Z = \dfrac{65-68}{3} = -1$

When $X = 71$, $Z = \dfrac{71-68}{3} = 1$

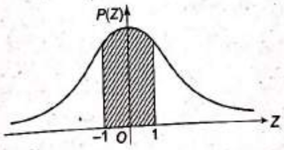


Fig. 5.30

$$\begin{aligned} P(65 \le X \le 71) &= P(-1 \le Z \le 1) \quad \text{(Fig. 5.30)} \\ &= P(-1 \le Z \le 0) + P(0 \le Z \le 1) \\ &= P(0 \le Z \le 1) + P(0 \le Z \le 1) \qquad \text{[By symmetry]} \\ &= 2P(0 \le Z \le 1) \\ &= 2(0.3413) \\ &= 0.6826 \end{aligned}$$

$$\begin{aligned} \text{Number of students with weights between 65 and 71 kg} &= N\,P(65 \le X \le 71) \\ &= 300(0.6826) \\ &= 204.78 \\ &\approx 205 \end{aligned}$$

Example 10

The mean yield for a one-acre plot is 662 *kg with an SD of* 32 *kg. Assuming normal distribution, how many one-acre plots in a batch of* 1000 *plots would you expect to have yields (i) over* 700 *kg? (ii) below* 650 *kg? (iii) What is the lowest yield of the best* 100 *plots?*

Solution

Let X be the random variable which denotes the yield for the one-acre plot.

$$\mu = 662, \quad \sigma = 32, \quad N = 1000$$

$$Z = \frac{X-\mu}{\sigma}$$

(i) When $X = 700$, $Z = \dfrac{700-662}{32} = 1.19$

$$\begin{aligned} P(X > 700) &= P(Z > 1.19) \quad \text{(Fig. 5.31)} \\ &= 0.5 - P(0 \le Z \le 1.19) \\ &= 0.5 - 0.3830 \\ &= 0.1170 \end{aligned}$$

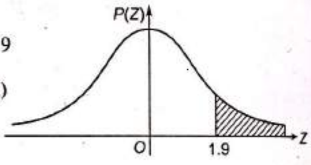


Fig. 5.31

$$\begin{aligned} \text{Expected number of plots with yields over 700 kg} &= N\,P(X > 700) \\ &= 1000\,(0.1170) \\ &= 117 \end{aligned}$$

(ii) When $X = 650$,

$$Z = \frac{650-662}{32} = -0.38$$

$$\begin{aligned} P(X < 650) &= P(Z < -0.38) \quad \text{(Fig. 5.32)} \\ &= P(Z > 0.38) \\ &\quad \text{[By symmetry]} \\ &= 0.5 - P(0 \le Z \le 0.38) \\ &= 0.5 - 0.1480 \\ &= 0.352 \end{aligned}$$

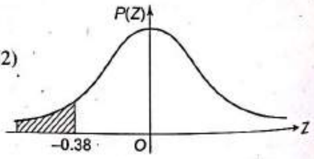


Fig. 5.32

$$\begin{aligned} \text{Expected number of plots with yields below 650 kg} &= N\,P(X < 650) \\ &= 1000(0.352) \\ &= 352 \end{aligned}$$

(iii) The lowest yield, say, x_1 of the best 100 plots is given by

$$P(X > x_1) = \frac{100}{1000} = 0.1$$

When $X = x_1$, $Z = \dfrac{x_1 - 662}{32} = z_1$

$$\begin{aligned} P(X > x_1) &= P(Z > z_1) \\ 0.1 &= 0.5 - P(0 \le Z \le z_1) \\ P(0 \le Z \le z_1) &= 0.4 \end{aligned}$$

$\therefore \quad z_1 = 1.2$ (approx.) [From normal table]

$$\frac{x_1 - 662}{32} = 1.28$$

$$x_1 = 702.96$$

Hence, the best 100 plots have yields over 702.96 kg.

Example 11

Assume that the mean height of Indian soldiers is 68.22 *inches with a variance of* 10.8 *inches. How many soldiers in a regiment of* 1000 *would you expect to be over* 6 *feet tall?*

Solution

Let X be the continuous random variable which denotes the heights of Indian soldiers.

$$\mu = 68.22, \quad \sigma^2 = 10.8, \quad N = 1000$$

$$\sigma = 3.29$$

$$Z = \frac{X-\mu}{\sigma}$$

When $X = 6$ feet $= 72$ inches,

$$Z = \frac{72 - 68.22}{3.29} = 1.15$$

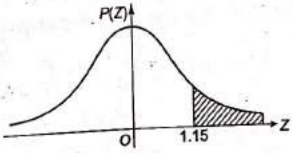


Fig. 5.33

$$\begin{aligned} P(X > 72) &= P(Z > 1.15) \quad \text{(Fig. 5.33)} \\ &= 0.5 - P(0 \le Z \le 1.15) \\ &= 0.5 - 0.3749 \\ &= 0.1251 \end{aligned}$$

Expected number of Indian soldiers having heights over 6 feet (72 inches)

$$\begin{aligned} &= N\,P(X > 72) \\ &= 1000(0.1251) \\ &= 125.1 \\ &\approx 125 \end{aligned}$$

Example 12

The marks obtained by students in a college are normally distributed with a mean of 65 and a variance of 25. If 3 students are selected at random from this college, what is the probability that at least one of them would have scored more than 75 marks?

Solution

Let X be the continuous random variable which denotes the marks of a student.

$$\mu = 65, \quad \sigma^2 = 25$$
$$\sigma = 5$$
$$Z = \frac{X-\mu}{\sigma}$$

When $X = 75$, $Z = \frac{75-65}{5} = 2$

$$P(X > 75) = P(Z > 2) \quad \text{(Fig. 5.34)}$$
$$= 0.5 - P(0 \le Z \le 2)$$
$$= 0.5 - 0.4772$$
$$= 0.0228$$

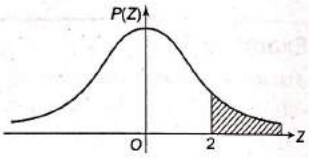


Fig. 5.34

If p is the probability of scoring more than 75 marks,

$$p = 0.0228, \quad q = 1 - p = 1 - 0.0228 = 0.9772$$

P(at least one student would have scored more than 75 marks)

$$= \sum_{x=1}^{3} {}^3C_x \, p^x \, q^{n-x}$$
$$= \sum_{x=1}^{3} {}^3C_x (0.0228)^x (0.9772)^{3-x}$$
$$= 0.0668$$

Example 13

Find the mean and standard deviation in which 7% of items are under 35 and 89% are under 63.

Solution

Let μ be the mean and σ be standard deviation of the normal curve.

$$P(X < 35) = 0.07$$
$$P(X < 63) = 0.89$$

$$P(X > 63) = 1 - P(X < 63) = 1 - 0.89 = 0.11$$
$$Z = \frac{X-\mu}{\sigma}$$

Since $P(X < 35) < 0.5$, the corresponding value of Z will be negative.

When $X = 35$, $Z = \frac{35-\mu}{\sigma} = -z_1$ (say)

Since $P(X < 63) > 0.5$, the corresponding value of Z will be positive.

When $X = 63$, $Z = \frac{63-\mu}{\sigma} = z_2$ (say)

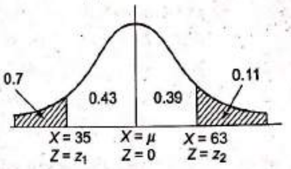


Fig. 5.35

From Fig. 5.35,

$$P(Z < -z_1) = 0.07$$
$$P(Z > z_2) = 0.11$$
$$P(0 < Z < z_1) = P(-z_1 < Z < 0)$$
$$= 0.5 - P(Z \le -z_1)$$
$$= 0.5 - 0.07$$
$$= 0.43$$
$$z_1 = 1.48 \quad \text{[From normal table]}$$
$$P(0 < Z < z_2) = 0.5 - P(Z \ge z_2)$$
$$= 0.5 - 0.11$$
$$= 0.39$$
$$z_2 = 1.23 \quad \text{[From normal table]}$$

Hence, $\frac{35-\mu}{\sigma} = -1.48$

$$-1.48\,\sigma + \mu = 35 \quad \text{...(1)}$$

and $\frac{63-\mu}{\sigma} = 1.23$

$$1.23\,\sigma + \mu = 63 \quad \text{...(2)}$$

Solving Eqs (1) and (2),

$$\mu = 50.29, \quad \sigma = 10.33$$

Example 14

In an examination, it is laid down that a student passes if he secures 40% or more. He is placed in the first, second, and third division according to whether he secures 60% or more marks, between 50% and 60% marks and between 40% and 50% marks respectively. He gets a distinction in case he secures 75% or more. It is noticed from the result that 10% of *the students failed in the examination, whereas 5% of them obtained distinction. Calculate the percentage of students placed in the second division. (Assume normal distribution of marks.)*

Solution

Let X be the random variable which denotes the marks of students in the examination. Let μ be the mean and σ be the standard deviation of the normal distribution of marks.

$$P(X<40)=0.10$$
$$P(X\geq 75)=0.05$$
$$P(X<75)=1-P(X\geq 75)=1-0.05=0.95$$
$$Z=\frac{X-\mu}{\sigma}$$

Since $P(X<40)<0.5$, the corresponding value of Z will be negative.

When $X=40$, $Z=\dfrac{40-\mu}{\sigma}=-z_1$ (say)

Since $P(X<75)<0.5$, the corresponding value of Z will be positive.

When $X=75$, $Z=\dfrac{75-\mu}{\sigma}=z_2$ (say)

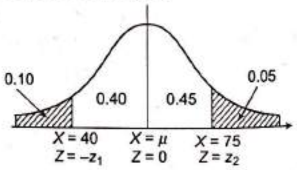


Fig. 5.36

From Fig. 5.36,

$$\begin{aligned}
P(Z<-z_1)&=0.10\\
P(Z>z_2)&=0.05\\
P(0<Z<z_1)&=P(-z_1<Z<0)\\
&=0.5-P(Z\leq -z_1)\\
&=0.5-0.10\\
&=0.40\\
z_1&=1.28 \quad \text{[From normal table]}\\
P(0<Z<z_2)&=0.5-P(Z\geq z_2)\\
&=0.5-0.05\\
&=0.45\\
z_2&=1.64 \quad \text{[From normal table]}
\end{aligned}$$

Hence, $\dfrac{40-\mu}{\sigma}=-1.28$

$$\mu-1.28\,\sigma=40 \qquad \text{...(1)}$$

and $\dfrac{75-\mu}{\sigma}=1.64$

$$\mu+1.64\,\sigma=75 \qquad \text{...(2)}$$

Solving Eqs (1) and (2),

$$\mu=55.34\approx 55$$
$$\sigma=11.98\approx 12$$

Probability that a student is placed in the second division is equal to the probability that his score lies between 50 and 60

When $X=50$, $Z=\dfrac{50-55}{12}=-0.42$

When $X=60$, $Z=\dfrac{60-55}{12}=0.42$

$$\begin{aligned}
P(50<X<60)&=P(-0.42<Z<0.42)\\
&=P(-0.42<Z<0)+P(0<Z<0.42)\\
&=P(0<Z<0.42)+P(0<Z<0.42) \quad \text{[By symmetry]}\\
&=2\,P(0<Z<0.42)\\
&=2\,(0.1628)\\
&=0.3256\\
&\approx 0.32
\end{aligned}$$

Hence, the percentage of students placed in the second division = 32%.

5.4.5 Fitting a Normal Distribution

Fitting a normal distribution or a normal curve to the data means to find the equation of the curve in the form $f(x)=\dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ which will be as close as possible to the points given. There are two purposes of fitting a normal curve:

(i) To judge the whether the normal curve is the best fit to the sample data.

(ii) To use the normal curve to estimate the characteristics of a population.

The area method for fitting a normal curve is given by the following steps:

(i) Find the mean μ and standard deviation σ for the given data if not given.

(ii) Write the class intervals and lower limits X of class intervals in two columns.

(iii) Find $Z=\dfrac{X-\mu}{\sigma}$ for each class interval.

(iv) Find the area corresponding to each Z from the normal table.

(v) Find the area under the normal curve between the successive values of Z. These are obtained by subtracting the successive areas when the corresponding Z's have the same sign and adding them when the corresponding Z's have the same sign and adding them when the corresponding Z's opposite sign. sign and adding them when the corresponding Z's

(vi) Find the expected frequencies by multiplying the relative frequencies by the number of observations.

Example 1

Fit a normal curve from the following distribution. It is given that the mean of the distribution is 43.7 and its standard distribution is 14.8.

Class interval	11–20	21–30	31–40	41–50	51–60	61–70	71–80
Frequency	20	28	40	60	32	20	8

Solution

$$\mu = 43.7, \quad \sigma = 14.8 \quad N = \Sigma f = 200$$

The series is converted into an inclusive series.

Class Interval	Lower class	$Z = \frac{X-\mu}{\sigma}$	Area from 0 to Z	Area in class Interval	Expected Frequencies
10.5–20.5	10.5	–2.24	0.4875	0.0457	9.14 ≈ 9
20.5–30.5	20.5	–1.57	0.4418	0.1285	25.7 ≈ 26
30.5–40.5	30.5	–0.89	0.3133	0.2262	45.24 ≈ 45
40.5–50.5	40.5	–0.22	0.0871	0.2643	52.86 ≈ 53
50.5–60.5	50.5	0.46	0.1772	0.1957	39.14 ≈ 39
60.5–70.5	60.5	1.14	0.3729	0.092	18.4 ≈ 18
70.5–80.5	70.5	1.81	0.4649	0.0287	5.74 ≈ 65
	80.5	2.49	0.4936		

Example 2

Fit a normal distribution to the following data:

X	125	135	145	155	165	175	185	195	205
Y	1	1	14	22	25	19	13	3	2

It is given that $\mu = 165.5$ and $\sigma = 15.26$.

Solution

$$\mu = 165.5, \quad \sigma = 15.26 \quad N = \Sigma f = 100$$

The data is first converted into class intervals with inclusive series.

Class Interval	Lower class	$Z = \frac{X-\mu}{\sigma}$	Area from 0 to Z	Area in class Interval	Expected Frequencies
120–130	120	–2.98	0.4986	0.0085	0.85 ≈ 1
130–140	130	–2.33	0.4901	0.0376	3.74 ≈ 4
140–150	140	–1.67	0.4525	0.1064	10.64 ≈ 11
150–160	150	–1.02	0.3461	0.2055	20.55 ≈ 21
160–170	160	–0.36	0.1406	0.2547	25.47 ≈ 25
170–180	170	0.29	0.1141	0.2148	21.48 ≈ 21
180–190	180	0.95	0.3289	0.1174	11.74 ≈ 12
190–200	190	1.61	0.4463	0.0418	4.18 ≈ 4
200–210	200	2.26	0.4881	0.0101	1.01 ≈ 1
210–220	210	2.92	0.4982		

EXERCISE 5.3

1. If X is normally distributed with a mean and standard deviation of 4, find (i) $P(5 \le X \le 10)$, (ii) $P(X \ge 15)$, (iii) $P(10 \le X \le 15)$, and (iv) $P(X \le 5)$.
[**Ans.:** (i) 0.3345 (ii) 0.003 (iii) 0.0638 (iv) 0.4013]

2. A normal distribution has a mean of 5 and a standard deviation of 3. What is the probability that the deviation from the mean of an item taken at random will be negative?
[**Ans.:** 0.0575]

3. If X is a normal variate with a mean of 30 and an SD of 6, find the value of $X = x_1$ such that $P(X \ge x_1) = 0.05$.
[**Ans.:** 39.84]

4. If X is a normal variate with a mean of 25 and SD of 5, find the value of $X = x_1$ such that $P(X \le x_1) = 0.01$.
[**Ans.:** 11.02]

5. The weights of 4000 students are found to be normally distributed with a mean of 50 kg and an SD of 5 kg. Find the probability that a student selected at random will have weight (i) less than 45 kg, and (ii) between 45 and 60 kg.
[**Ans.:** (i) 0.1587 (ii) 0.8185]

6. The daily sales of a firm are normally distributed with a mean of ₹ 8000 and a variance of ₹ 10000. (i) What is the probability that on a certain

day the sales will be less than ₹ 8210? (ii) What is the percentage of days on which the sales will be between ₹ 8100 and ₹ 8200?

[**Ans.:** (i) 0.482 (ii) 14%]

7. The mean height of Indian soldiers is 68.22″ with a variance of 10.8″. Find the expected number of soldiers in a regiment of 1000 whose height will be more than 6 feet.

[**Ans.:** 125]

8. The life of army shoes is normally distributed with a mean of 8 months and a standard deviation of 2 months. If 5000 pairs are issued, how many pairs would be expected to need replacement after 12 months?

[**Ans.:** 2386]

9. In an intelligence test administered to 1000 students, the average was 42 and the standard deviation was 24. Find the number of students (i) exceeding 50, (ii) between 30 and 54, and (iii) the least score of top 1000 students.

[**Ans.:** (i) 129 (ii) 383 (iii) 72.72]

10. In a test of 2000 electric bulbs, it was found that the life of a particular make was normally distributed with an average of life of 2040 hours and a standard deviation of 60 hours. Estimate the number of bulbs likely to burn for (i) more than 2150 hours, and (ii) less than 1950 hours.

[**Ans.:** (i) 67 (ii) 184]

11. The marks of 1000 students of a university are found to be normally distributed with a mean of 70 and a standard of deviation 5. Estimate the number of students whose marks will be (i) between 60 and 75, (ii) more than 75, and (iii) less than 68.

[**Ans.:** (i) 910 (ii) 23 (iii) 37]

12. In a normal distribution, 31% items are under 45 and 8% are over 64. Find the mean and standard deviation. Find also, the percentage of items lying between 30 and 75.

[**Ans.:** 50, 10, 0.957]

13. Of a large group of men, 5% are under 60 inches in height and 40% are between 60 and 65 inches. Assuming a normal distribution, find the mean and standard deviation of distribution.

[**Ans.:** 65.42, 3.27]

14. The marks obtained by students in an examination follow a normal distribution. If 30% of the students got marks below 35 and 10% got marks above 60, find the mean and percentage of students who got marks between 40 and 50.

[**Ans.:** 42.23, 13.88, 28%]

15. Fit a normal distribution to the following data:

Class	60–65	65–70	70–75	75–80	80–85	85–90	90–95	95–100
Frequency	3	21	150	335	326	135	26	4

[**Ans.:** Expected frequency: 3, 31, 148, 322, 319, 144, 30, 3]

5.5 EXPONENTIAL DISTRIBUTION

A continuous random variable X is said to follow exponential distribution if its probability function is given by

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0$$
$$= 0 \quad , \quad x \le 0$$

where $\lambda > 0$ is called the rate of the distribution.

5.5.1 Memoryless Property of the Exponential Distribution

The exponential distribution has the memoryless (forgetfulness) property. This property indicates that the distribution is independent of its part, that means future happening of an event has no relation to whether or not this even has happened in the past. This property is as follows:

If X is exponentially distributed, and s, t are two positive real numbers then

$$P[(X > s+t)/(X > s)] = P(X > t)$$

Proof: $P[(X > s+t)/(X > s)] = \dfrac{P[(X > s+t) \cap (X > s)]}{P(X > s)}$

[using conditional probability]

$$= \frac{P(X > s+t)}{P(X > s)}$$

$$= \frac{\int_{s+t}^{\infty} \lambda e^{-\lambda x} dx}{\int_{s}^{\infty} \lambda e^{-\lambda x} dx}$$

$$= \frac{\lambda \left| \frac{e^{-\lambda x}}{-\lambda} \right|_{s+t}^{\infty}}{\lambda \left| \frac{e^{-\lambda x}}{-\lambda} \right|_{s}^{\infty}}$$

$$= \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}}$$

$$= e^{-\lambda t} \qquad \text{...(5.5)}$$

$$P(X > t) = \int_t^{\infty} \lambda e^{-\lambda x} \mathrm{d}x$$

$$= \lambda \left| \frac{e^{-\lambda x}}{-\lambda} \right|_t^{\infty}$$

$$= e^{-\lambda t} \qquad \text{...(5.6)}$$

From Eq. (5.5) and Eq. (5.6),

$$P[(X > s+t)/(X > s)] = P(X > t), \qquad \text{for } s, t > 0$$

5.5.2 Constants of the Exponential Distribution

1. Mean of the Exponential Distribution

$$E(X) = \int_{-\infty}^{\infty} x\, f(x) \mathrm{d}x$$

$$= \int_0^{\infty} x\, \lambda e^{-\lambda x} \mathrm{d}x$$

$$= \lambda \left| x \cdot \frac{e^{-\lambda x}}{-\lambda} - 1 \cdot \frac{e^{-\lambda x}}{\lambda^2} \right|_0^{\infty}$$

$$= \lambda \cdot \frac{1}{\lambda^2}$$

$$= \frac{1}{\lambda}$$

2. Variance of the Exponential Distribution

$$\text{Var}(X) = E(X^2) - [E(X)]^2 \qquad \text{...(5.7)}$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x) \mathrm{d}x$$

$$= \int_0^{\infty} x^2 \lambda e^{-\lambda x} \mathrm{d}x$$

$$= \lambda \left| x^2 \frac{e^{-\lambda x}}{-\lambda} - 2x \frac{e^{-\lambda x}}{\lambda^2} + 2 \frac{e^{-\lambda x}}{-\lambda^3} \right|_0^{\infty}$$

$$= \lambda \left(\frac{2}{\lambda^3} \right)$$

$$= \frac{2}{\lambda^2}$$

Substituting in Eq (5.7),

$$\text{Var}(X) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2} \qquad \left[\because \mu = \frac{1}{\lambda} \right]$$

3. Standard Deviation of the Exponential Distribution

$$\text{SD} = \sqrt{\text{Var}(X)} = \sqrt{\frac{1}{\lambda^2}} = \frac{1}{\lambda}$$

4. Mode of the Exponential Distribution

Mode is the value of x for which $f(x)$ is maximum.

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0$$
$$= 0 \quad , \quad x \le 0$$

$f(x)$ will be maximum when $e^{-\lambda x}$ is maximum.

Maximum value of $e^{-\lambda x} = 1$, which is at $x = 0$.

Hence, $x = 0$ is the mode of the exponential distribution

5. Median of the Exponential Distribution

If M is the median of the exponential distribution,

$$\int_{-\infty}^{M} f(x)\, \mathrm{d}x = \frac{1}{2}$$

$$\int_0^{M} \lambda e^{-\lambda x}\, \mathrm{d}x = \frac{1}{2}$$

$$\lambda \left| \frac{e^{-\lambda x}}{-\lambda} \right|_0^{M} = \frac{1}{2}$$

$$-(e^{-\lambda M} - 1) = \frac{1}{2}$$

$$-e^{-\lambda M} = \frac{1}{2} - 1 = -\frac{1}{2}$$

$$e^{-\lambda M} = \frac{1}{2}$$

$$-\lambda M \log e = \log\frac{1}{2} = -\log 2$$

$$\lambda M = \log 2$$

$$M = \frac{1}{\lambda}\log 2$$

Example 1

Let X be a random variable with pdf

$$f(x) = \begin{cases} \frac{1}{5}e^{-\frac{x}{5}} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Find (i) $P(X > 5)$ (ii) $P(3 \le X \le 6)$ (iii) mean (iv) variance.

Solution

$$\lambda = \frac{1}{5}$$

(i) $P(X > 5) = \int_5^\infty f(x)\,dx$

$$= \int_5^\infty \frac{1}{5}e^{-\frac{x}{5}}\,dx$$

$$= \frac{1}{5}\left|\frac{e^{-\frac{x}{5}}}{-\frac{1}{5}}\right|_5^\infty$$

$$= -\left|e^{-\frac{x}{5}}\right|_5^\infty$$

$$= -(e^{-\infty} - e^{-1})$$

$$= e^{-1}$$

$$= 0.3679$$

(ii) $P(3 \le X \le 6) = \int_3^6 f(x)\,dx$

$$= \int_3^6 \frac{1}{5}e^{-\frac{x}{5}}\,dx$$

$$= \frac{1}{5}\left|\frac{e^{-\frac{x}{5}}}{-\frac{1}{5}}\right|_3^6$$

$$= -\left|e^{-\frac{x}{5}}\right|_3^6$$

$$= -\left(e^{-\frac{6}{5}} - e^{-\frac{3}{5}}\right)$$

$$= e^{-\frac{3}{5}} - e^{-\frac{6}{5}}$$

$$= 0.2476$$

(iii) Mean $\mu = \frac{1}{\lambda} = \frac{1}{\left(\frac{1}{5}\right)} = 5$

(iv) Variance $= \text{Var}(X) = \frac{1}{\lambda^2} = \frac{1}{\left(\frac{1}{5}\right)^2} = 25$

Example 2

A random variable has pdf $f(x) = ce^{-2x}$ for $x > 0$. Find (i) $P(X > 2)$ (ii) $P\left(X < \frac{1}{c}\right)$.

Solution

Since $f(x)$ is a probability density function,

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$\int_0^\infty ce^{-2x}\,dx = 1$$

$$\left|\frac{ce^{-2x}}{-2}\right|_0^\infty = 1$$

$$-\frac{c}{2}\left|e^{-2x}\right|_0^\infty = 1$$

$$-\frac{c}{2}(e^{-\infty} - e^0) = 1$$

$$\frac{c}{2} = 1$$

$$c = 2$$

$$\therefore f(x) = 2e^{-2x}, \qquad x > 0$$

(i) $P(X > 2) = \int_2^\infty f(x)\,dx$

$$= \int_2^\infty 2e^{-2x}\,dx$$

$$= 2\left|\frac{e^{-2x}}{-2}\right|_2^\infty$$

$$= -\left|e^{-2x}\right|_2^\infty$$

$$= -(e^{-\infty} - e^{-4})$$

$$= e^{-4}$$

$$= 0.0183$$

(ii) $P\left(X < \frac{1}{c}\right) = P\left(X < \frac{1}{2}\right)$

$$= \int_0^{\frac{1}{2}} f(x)\,dx$$

$$= \int_0^{\frac{1}{2}} 2e^{-2x}\,dx$$

$$= 2\left|\frac{e^{-2x}}{-2}\right|_0^{\frac{1}{2}}$$

$$= -\left|e^{-2x}\right|_0^{\frac{1}{2}}$$

$$= -(e^{-1} - e^0)$$

$$= -e^{-1} + 1$$

$$= 0.6321$$

Example 3

If X is random variable which follows an exponential distribution with parameter λ with $P(X \le 1) = P(X > 1)$, find $\text{Var}(X)$.

Solution

Since X is random variable which follows an exponential distribution,

$$f(x) = \lambda e^{-\lambda x}, \quad x \ge 0$$

$$P(X \le 1) = P(X > 1)$$

$$1 - P(X > 1) = P(X > 1)$$

$$2P(X > 1) = 1$$

$$P(X > 1) = \frac{1}{2}$$

$$\int_1^\infty f(x)\,dx = \frac{1}{2}$$

$$\int_1^\infty \lambda e^{-\lambda x}\,dx = \frac{1}{2}$$

$$\lambda\left|\frac{e^{-\lambda x}}{-\lambda}\right|_1^\infty = \frac{1}{2}$$

$$-\left|e^{-\lambda x}\right|_1^\infty = \frac{1}{2}$$

$$-(e^{-\infty} - e^{-\lambda}) = \frac{1}{2}$$

$$e^{-\lambda} = \frac{1}{2}$$

$$\frac{1}{e^\lambda} = \frac{1}{2}$$

$$e^\lambda = 2$$

$$\lambda = \log_e 2$$

$$\text{Var}(X) = \frac{1}{\lambda^2} = \frac{1}{(\log_e 2)^2}$$

Example 4

If X is a exponentially distributed random variable with parameter λ, find the value of k such that $\frac{P(X > k)}{P(X \le k)} = a$.

Solution

$$\frac{P(X>k)}{P(X\le k)}=a$$

$$\frac{P(X>k)}{1-P(X>k)}=a$$

$$P(X>k)=a\left[1-P(X>k)\right]$$

$$P(X>k)(1+a)=a$$

$$P(X>k)=\frac{a}{1+a}$$

$$\int_k^\infty f(x)\,\mathrm{d}x=\frac{a}{1+a}$$

$$\int_k^\infty \lambda e^{-\lambda x}\,\mathrm{d}x=\frac{a}{1+a}$$

$$\lambda\left|\frac{e^{-\lambda x}}{-\lambda}\right|_k^\infty=\frac{a}{1+a}$$

$$-\left|e^{-\lambda x}\right|_k^\infty=\frac{a}{1+a}$$

$$-(e^{-\infty}-e^{-\lambda k})=\frac{a}{1+a}$$

$$e^{-\lambda k}=\frac{a}{1+a}$$

$$\frac{1}{e^{\lambda k}}=\frac{a}{1+a}$$

$$e^{\lambda k}=\frac{1+a}{a}$$

$$\lambda k=\log\left(\frac{1+a}{a}\right)$$

$$k=\frac{1}{\lambda}\log\left(\frac{1+a}{a}\right)$$

Example 5

If the density function of a continuous random variable X is $f(x) = ce^{-b(x-a)}$, $a \le x$ where a, b, c are constants. Show that $b = c = \frac{1}{\sigma}$ and $a = \mu - \sigma$, where $\mu = E(X)$ and $\sigma^2 = \mathrm{Var}(X)$.

Solution

Since $f(x)$ is a density function,

$$\int_{-\infty}^\infty f(x)\,\mathrm{d}x=1$$

$$\int_a^\infty ce^{-b(x-a)}\,\mathrm{d}x=1$$

$$c\left|\frac{e^{-b(x-a)}}{-b}\right|_a^\infty=1$$

$$-\frac{c}{b}\left|e^{-b(x-a)}\right|_a^\infty=1$$

$$-\frac{c}{b}(e^{-\infty}-e^{0})=1$$

$$\frac{c}{b}=1$$

$$b=c \qquad \text{...(1)}$$

$$\mu=E(X)=\int_a^\infty bxe^{-b(x-a)}\,\mathrm{d}x$$

$$=be^{ab}\left|x\left(\frac{e^{-bx}}{-b}\right)-\frac{e^{-bx}}{b^2}\right|_a^\infty$$

$$=be^{ab}\left(\frac{a}{b}e^{-ab}+\frac{1}{b^2}e^{-ab}\right)$$

$$=a+\frac{1}{b} \qquad \text{...(2)}$$

$$E(X^2)=\int_a^\infty bx^2e^{-b(x-a)}\,\mathrm{d}x$$

$$=be^{ab}\left|x^2\left(\frac{e^{-bx}}{-b}\right)-2x\left(\frac{e^{-bx}}{-b^2}\right)+2\left(\frac{e^{-bx}}{-b^3}\right)\right|_a^\infty$$

$$=b\left(\frac{a^2}{b}+\frac{2a}{b^2}+\frac{2}{b^3}\right)$$

$$=\frac{1}{b^2}(a^2b^2+2ab+2)$$

$$\begin{aligned}\text{Var}(X) &= E(X^2)-[E(X)]^2\\ \sigma^2 &= \frac{1}{b^2}(a^2b^2+2ab+2)-\left(a^2+\frac{2a}{b}+\frac{1}{b^2}\right)\\ &= \frac{1}{b^2}\\ \sigma &= \frac{1}{b} \qquad \text{...(3)}\end{aligned}$$

From Eq. (1) and (3),

$$b=c=\frac{1}{\sigma}$$

Subtracting Eq. (3) from Eq. (2),

$$\mu-\sigma=a$$

$$\therefore \quad a=\mu-\sigma$$

Example 6

The mileage which car owners get with a certain kind of radial tire is a random variable having an exponential distribution with mean 4000 *km. Find the probabilities that one of these tires will last (i) at least* 2000 *km (ii) at most* 3000 *km.*

Solution

Let X be the random variable which denotes the mileage obtained with the tire.

$$\begin{aligned}\text{Mean } \mu &= \frac{1}{\lambda}=4000 \text{ km}\\ f(x) &= \lambda e^{-\lambda x}, \qquad x>0\\ &= \frac{1}{4000}e^{-\frac{1}{4000}x}, \qquad x>0\end{aligned}$$

(i) $$\begin{aligned}P(X\ge 2000) &= \int_{2000}^{\infty} f(x)\,\mathrm{d}x\\ &= \int_{2000}^{\infty}\frac{1}{4000}e^{-\frac{1}{4000}x}\,\mathrm{d}x\\ &= \frac{1}{4000}\left|\frac{e^{-\frac{1}{4000}x}}{-\frac{1}{4000}}\right|_{2000}^{\infty}\end{aligned}$$

$$\begin{aligned}&= -\left|e^{-\frac{1}{4000}x}\right|_{2000}^{\infty}\\ &= -(e^{-\infty}-e^{-0.5})\\ &= e^{-0.5}\\ &= 0.6065\end{aligned}$$

(ii) $$\begin{aligned}P(X\le 3000) &= \int_0^{3000} f(x)\,\mathrm{d}x\\ &= \int_0^{3000}\frac{1}{4000}e^{-\frac{1}{4000}x}\,\mathrm{d}x\\ &= \frac{1}{4000}\left|\frac{e^{-\frac{1}{4000}x}}{-\frac{1}{4000}}\right|_0^{3000}\\ &= -\left|e^{-\frac{1}{4000}x}\right|_0^{3000}\\ &= -(e^{-0.75}-e^{0})\\ &= -e^{-0.75}+1\\ &= 0.5270\end{aligned}$$

Example 7

If the number of kilometers that a car can run before its battery wears out is exponentially distributed with an average value of 10000 *km and if the owner desires to take a* 5000 *km trip, what is the probability that he will be able to complete his trip without having to replace the car battery. Assume that the car has been used for same time.*

Solution

Let X be the random variable which denotes the number of kilometers that a car can run before its battery wears out.

$$\begin{aligned}\text{Mean } \mu &= \frac{1}{\lambda}=10000\\ f(x) &= \lambda e^{-\lambda x}, \qquad x>0\\ &= \frac{1}{1000}e^{-\frac{1}{10000}x}, x>0\end{aligned}$$

$$
\begin{aligned}
P(X>5000) &= \int_{5000}^{\infty} f(x)\,\mathrm{d}x \\
&= \int_{5000}^{\infty} \frac{1}{10000} e^{-\frac{1}{10000}x}\,\mathrm{d}x \\
&= \frac{1}{10000}\left|\frac{e^{-\frac{1}{10000}x}}{-\frac{1}{10000}}\right|_{5000}^{\infty} \\
&= -\left|e^{-\frac{1}{10000}x}\right|_{5000}^{\infty} \\
&= -(e^{-\infty}-e^{-0.5}) \\
&= e^{-0.5} \\
&= 0.6065
\end{aligned}
$$

Example 8

The average time it takes to serve a customer at a petrol pump is 6 minutes. The service time follows exponential distribution. Calculate the probability that

(i) *A customer will take less than 2 minutes to complete the service.*
(ii) *A customer will take between 4 and 5 minutes to get the service.*
(iii) *A customer will take more than 10 minutes for his service.*

Solution

Let X be the random variable which denotes the service time.

$$
\begin{aligned}
\text{Mean } \mu &= \frac{1}{\lambda} = 6 \\
f(x) &= \lambda e^{-\lambda x},\ x>0 \\
&= \frac{1}{6} e^{-\frac{1}{6}x},\ x>0
\end{aligned}
$$

(i) $P(X<2) = \int_0^2 f(x)\,\mathrm{d}x$

$$
\begin{aligned}
&= \int_0^2 \frac{1}{6} e^{-\frac{1}{6}x}\,\mathrm{d}x \\
&= \frac{1}{6}\left|\frac{e^{-\frac{1}{6}x}}{-\frac{1}{6}}\right|_0^2 \\
&= -\left|e^{-\frac{1}{6}x}\right|_0^2 \\
&= -(e^{-\frac{1}{3}}-e^0) \\
&= -e^{-\frac{1}{3}}+1 \\
&= 0.2835
\end{aligned}
$$

(ii) $P(4<X<5) = \int_4^5 f(x)\,\mathrm{d}x$

$$
\begin{aligned}
&= \int_4^5 \frac{1}{6} e^{-\frac{1}{6}x}\,\mathrm{d}x \\
&= \frac{1}{6}\left|\frac{e^{-\frac{1}{6}x}}{-\frac{1}{6}}\right|_4^5 \\
&= -\left|e^{-\frac{1}{6}x}\right|_4^5 \\
&= -\left(e^{-\frac{5}{6}}-e^{-\frac{2}{3}}\right) \\
&= 0.0788
\end{aligned}
$$

(iii) $P(X>10) = \int_{10}^{\infty} f(x)\,\mathrm{d}x$

$$
\begin{aligned}
&= \int_{10}^{\infty} \frac{1}{6} e^{-\frac{1}{6}x}\,\mathrm{d}x \\
&= \frac{1}{6}\left|\frac{e^{-\frac{1}{6}x}}{-\frac{1}{6}}\right|_{10}^{\infty} \\
&= -\left|e^{-\frac{1}{6}x}\right|_{10}^{\infty}
\end{aligned}
$$

$$= -\left(e^{-\infty} - e^{-\frac{10}{6}}\right)$$

$$= e^{-\frac{10}{6}}$$

$$= 0.1889$$

Example 9

The length of time X to complete a job is exponentially distributed with $E(X) = \mu = \frac{1}{\lambda} = 10$ *hours. (i) Compute the probability of job completion between two consecutive jobs exceeding 20 hours. (ii) The cost of job completion is given by* $C = 4 + 2X + 2X^2$*. Find the expected value of C.*

Solution

Let X be a random variable which denotes the length of time to complete a job.

$$E(X) = \mu = \frac{1}{\lambda} = 10$$

$$f(x) = \lambda e^{-\lambda x}$$

$$= \frac{1}{10} e^{-\frac{1}{10}x}$$

(i) $P(X > 20) = \int_{20}^{\infty} f(x)\,\mathrm{d}x$

$$= \int_{20}^{\infty} \frac{1}{10} e^{-\frac{1}{10}x}\,\mathrm{d}x$$

$$= \frac{1}{10}\left|\frac{e^{-\frac{1}{10}x}}{-\frac{1}{10}}\right|_{20}^{\infty}$$

$$= -\left|e^{-\frac{1}{10}x}\right|_{20}^{\infty}$$

$$= -(e^{-\infty} - e^{-2})$$

$$= e^{-2}$$

$$= 0.1353$$

(ii) For an exponential random variable,

$$E(X) = \mu = \frac{1}{\lambda} = 10$$

$$\text{Var}\,(X) = \frac{1}{\lambda^2}$$

$$\text{Var}\,(X) = E(X^2) - \mu^2$$

$$E(X^2) = \text{Var}\,(X) + \mu^2$$

$$= \frac{1}{\lambda^2} + \frac{1}{\lambda^2}$$

$$= \frac{2}{\lambda^2}$$

$$= 200$$

$$E(C) = E(4 + 2X + 2X^2)$$

$$= E(4) + 2E(X) + 2E(X^2)$$

$$= 4 + 2(10) + 2(200)$$

$$= 424$$

Example 10

The time (in hours) required to repair a machine is exponentially distributed with parameter $\lambda = \frac{1}{2}$.

(i) *What is the probability that the repair time exceeds 2 hours?*

(ii) *What is the conditional probability that a repair takes at least 11 hours given that its direction exceeds 8 hours?*

Solution

Let X be the random variable which denotes the time to repair the machine.

$$\lambda = \frac{1}{2}$$

$$f(x) = \lambda e^{-\lambda x}, \qquad x > 0$$

$$= \frac{1}{2} e^{-\frac{1}{2}x}, \qquad x > 0$$

(i) $$P(X>2)=\int_2^\infty f(x)\,\mathrm{d}x$$

$$=\int_2^\infty \frac{1}{2}e^{-\frac{1}{2}x}\,\mathrm{d}x$$

$$=\frac{1}{2}\left|\frac{e^{-\frac{1}{2}x}}{-\frac{1}{2}}\right|_2^\infty$$

$$=-\left|e^{-\frac{1}{2}x}\right|_2^\infty$$

$$=-(e^{-\infty}-e^{-1})$$

$$=e^{-1}$$

$$=0.3679$$

(ii) $P(X\geq 11/X>9)=P(X>3)$ (By the memoryless property)

$$=\int_3^\infty f(x)\,\mathrm{d}x$$

$$=\int_3^\infty \frac{1}{2}e^{-\frac{1}{2}x}\,\mathrm{d}x$$

$$=\frac{1}{2}\left|\frac{e^{-\frac{1}{2}x}}{-\frac{1}{2}}\right|_3^\infty$$

$$=-\left|e^{-\frac{1}{2}x}\right|_3^\infty$$

$$=-(e^{-\infty}-e^{-1.5})$$

$$=e^{-1.5}$$

$$=0.2231$$

Example 11

The daily consumption of milk in excess of 20000 *gallons is approximately exponentially distributed with* $\lambda=\frac{1}{3000}$. *The city has a daily stock of* 35000 *gallons. What is the probability that of* 2 *days selected at random, the stock is insufficient for both the days.*

Solution

Let Y be a random variable which denotes the daily consumption of milk consumed in a day. The random variable $X=Y-20000$ has an exponential distribution.

$$\lambda=\frac{1}{3000}$$

$$f(x)=\lambda e^{-\lambda x},\qquad x>0$$

$$=\frac{1}{3000}e^{-\frac{1}{3000}x},\qquad x>0$$

Probability that the stock is insufficient for both days

$$P(Y>35000)=P(X>15000)$$

$$=\int_{15000}^\infty f(x)\,\mathrm{d}x$$

$$=\int_{15000}^\infty \frac{1}{3000}e^{-\frac{1}{3000}x}\,\mathrm{d}x$$

$$=\frac{1}{3000}\left|\frac{e^{-\frac{1}{3000}x}}{-\frac{1}{3000}}\right|_{15000}^\infty$$

$$=-\left|e^{-\frac{1}{3000}x}\right|_{15000}^\infty$$

$$=-(e^{-\infty}-e^{-5})$$

$$=e^{-5}$$

$$=0.0067$$

EXERCISE 5.4

1. If X is exponentially distributed, prove that probability that X exceeds its expected value is less than 0.5.
2. The amount of time that a watch will run without having to be reset is a random variable having an exponential distribution with mean 120 days. Find the probability that such a watch will
   (a) have to be set in less than 24 days.
   (b) not have to be reset in at least 180 days.

[Ans.: (a) 0.1813, (b) 0.2231]

3. The length of the shower on a tropical island during rainy season has an exponential distribution with parameter 2, time being measured in minutes. What is the probability that a shower will last more than 3 minutes? If a shower has already lasted for 2 minutes, what is the probability that it will last for at least one more minute?

[Ans.: (a) 0.0025, (b) 0.1353]

4. If X is exponentially distributed with parameter λ, find the value of k such that $P(X > k)/P(X \le k) = a$.

$$\left[\textbf{Ans.: } \lambda^{-1}\log\left(1+\frac{1}{a}\right)\right]$$

5. The life length X of an electronic component follows an exponential distribution. These are 2 processes by which the component may be manufactured. The expected life length of the component is 100 hrs if process I is used to manufacture, while it is 150 hrs if process II is used. The cost of manufacturing a single component by process I is ₹10, while is ₹20 for process II. Moreover, if the component lasts less than the guaranteed life of 200 hrs, a loss of ₹50 is to be borne by the manufacturer. Which process is advantageous to the manufacturer?

[Ans.: Process I is advantageous to the manufacturer]

6. The life of an electronic component follows exponential distribution with a mean of 4 years. The manufacturer of this component gives a replacement warranty of 3 years.
   (a) What proportion of components will be replaced in the period of warranty?
   (b) What is the probability that a randomly selected component will have life within two standard deviations of the mean life?

[Ans.: (a) 0.5276, (b) 0.9502]

5.6 GAMMA DISTRIBUTION

A continuous random variable X is said to follow exponential distribution if its probability function is given by

$$f(x) = \frac{\lambda^r}{\overline{|r}} x^{r-1} e^{-\lambda x}, \quad x > 0$$
$$= 0 \quad , \quad x \le 0$$

5.6.1 Constants of the Gamma Distribution

1. Mean of the Gamma Distribution

$$E(X) = \int_{-\infty}^{\infty} x f(x)\mathrm{d}x$$
$$= \int_0^{\infty} x \frac{\lambda^r}{\overline{|r}} x^{r-1} e^{-\lambda x}\mathrm{d}x$$
$$= \frac{\lambda^r}{\overline{|r}} \int_0^{\infty} e^{-\lambda x} x^r \mathrm{d}x$$
$$= \frac{\lambda^r}{\overline{|r}} \frac{\overline{|r+1}}{\lambda^{r+1}} \qquad \left[\because \int_0^{\infty} e^{-kx} x^{n-1}\mathrm{d}x = \frac{\overline{|n}}{k^n}\right]$$
$$= \frac{\lambda^r r \overline{|r}}{\overline{|r}\cdot\lambda^{r+1}}$$
$$= \frac{r}{\lambda}$$

2. Variance of the Gamma Distribution

$$\mathrm{Var}(X) = E(X^2) - [E(X)]^2 \qquad \text{...(5.8)}$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\mathrm{d}x$$
$$= \int_0^{\infty} x^2 \frac{\lambda^r}{\overline{|r}} x^{r-1} e^{-\lambda x}\mathrm{d}x$$
$$= \frac{\lambda^r}{\overline{|r}} \int_0^{\infty} e^{-\lambda x} x^{r+1}\mathrm{d}x$$
$$= \frac{\lambda^r}{\overline{|r}} \frac{\overline{|r+2}}{\lambda^{r+2}} \qquad \left[\because \int_0^{\infty} e^{-kx} x^{n-1}\mathrm{d}x = \frac{\overline{|n}}{k^n}\right]$$
$$= \frac{(r+1) r \overline{|r}}{\overline{|r}\cdot\lambda^2}$$
$$= \frac{r^2 + r}{\lambda^2}$$

Substituting in Eq. (5.8),

$$\mathrm{Var}(X) = \frac{r^2 + r}{\lambda^2} - \frac{r^2}{\lambda^2}$$
$$= \frac{r}{\lambda^2}$$

3. Standard Deviation of the Gamma Distribution

$$\text{SD} = \sqrt{\text{Var}(X)} = \sqrt{\frac{r}{\lambda^2}} = \frac{\sqrt{r}}{\lambda}$$

4. Mode of the Gamma Distribution

Mode is the value of x for which $f(x)$ is maximum.

$$\begin{aligned} f(x) &= \frac{\lambda^r}{\overline{|r}} x^{r-1} e^{-\lambda x}, && x > 0 \\ &= 0 \quad , && x \le 0 \end{aligned}$$

Differentiating w.r.t. x,

$$\begin{aligned} f'(x) &= \frac{\lambda^r}{\overline{|r}} \left[(r-1)x^{r-2}e^{-\lambda x} + x^{r-1}e^{-\lambda x}(-\lambda) \right] \\ &= \frac{\lambda^r}{\overline{|r}} x^{r-2} e^{-\lambda x} \left[(r-1) - \lambda x \right] \end{aligned}$$

For maximum value of $f(x)$,

$$\begin{aligned} f'(x) &= 0 \\ (r-1) - \lambda x &= 0 \\ x &= \frac{r-1}{\lambda} \end{aligned}$$

Differentiating $f'(x)$ w.r.t. x,

$$\begin{aligned} f''(x) &= \frac{\lambda^r}{\overline{|r}} \Big[(r-2)x^{r-3}e^{-\lambda x}(r-1-\lambda x) \\ &\qquad + x^{r-2}e^{-\lambda x}(-\lambda)(r-1-\lambda x) + x^{r-2}e^{-\lambda x}(-\lambda) \Big] \\ &= \frac{\lambda^r}{\overline{|r}} x^{r-3} e^{-\lambda x} \left[(r-2)(r-1-\lambda x) - \lambda x(r-1-\lambda x) - \lambda x \right] \end{aligned}$$

Putting $x = \frac{r-1}{\lambda}$,

$$\begin{aligned} f''(x) &= \frac{\lambda^r}{\overline{|r}} x^{r-3} e^{-\lambda x} \left[(r-2)(r-1-r+1) - \lambda x(r-1-r+1) - (r-1) \right] \\ &= \frac{\lambda^r}{\overline{|r}} x^{r-3} e^{-\lambda x} (1-r) \end{aligned}$$

$f(x)$ is maximum when $x = \frac{r-1}{\lambda}$, if $f''(x) < 0$,

$$\begin{aligned} f''(x) < 0 \text{ if } 1 - r &< 0 \\ 1 &< r \\ \text{or} \quad r &> 1 \end{aligned}$$

Hence, $x = \frac{r-1}{\lambda}$ is the mode of the gamma distribution for $r > 1$.

Example 1

Given a Gamma random variable X with r = 3 and λ = 2. Compute E(X), Var(X) *and P(X ≤ 1.5 years).*

Solution

$$\lambda = 2, \quad r = 3$$

$$f(x) = \frac{\lambda^r}{\overline{|r}} x^{r-1} e^{-\lambda x}, \quad x > 0$$

(a) $E(X) = \frac{r}{\lambda} = \frac{3}{2} = 1.5$ years

(b) $\text{Var}(X) = \frac{r}{\lambda^2} = \frac{3}{(2)^2} = 0.75$

(c) $P(X \le 1.5 \text{ years}) = \int_0^{1.5} f(x)\,dx$

$$\begin{aligned} &= \int_0^{1.5} \frac{2^3}{\overline{|3}} x^2 e^{-2x}\,dx \\ &= 4 \left| x^2\left(\frac{e^{-2x}}{-2}\right) - 2x\left(\frac{e^{-2x}}{4}\right) + 2\left(\frac{e^{-2x}}{-8}\right) \right|_0^{1.5} \\ &= 4\left[(1.5)^2\left(\frac{e^{-3}}{-2}\right) - 2(1.5)\left(\frac{e^{-3}}{4}\right) + 2\left(\frac{e^{-3}}{-8}\right) + \frac{1}{4} \right] \\ &= 0.5768 \end{aligned}$$

Example 2

The daily consumption of milk in a city, in excess 20000 litres, is approximately distributed as a Gamma variate with parameters $\lambda = \frac{1}{10000}$

and r = 2. The city has a daily stock of 30000 litres. What is the probability that the stock is insufficient on a particular day?

Solution

Let Y be the random variable which denotes the daily consumption of milk (in litres) in a city. The random variable $X = Y - 20000$ has a gamma distribution.

$$\lambda = \frac{1}{10000}, r = 2$$

$$\begin{aligned} f(x) &= \frac{\lambda^r}{\lceil r} x^{r-1} e^{-\lambda x}, \ x > 0 \\ &= \frac{\left(\frac{1}{10000}\right)^2}{\lceil 2} x^{2-1} e^{-\frac{1}{10000}x} \\ &= \frac{x e^{-\frac{1}{1000}x}}{(10000)^2} \end{aligned}$$

Probability that the stock is insufficient on a particular day

$$\begin{aligned} P(Y > 30000) &= P(X > 10000) \\ &= \int_{10000}^{\infty} f(x)\mathrm{d}x \\ &= \int_{1000}^{\infty} \frac{x e^{-\frac{1}{1000}x}}{(10000)^2}\mathrm{d}x \\ &= \frac{1}{10^8}\int_{10^4}^{\infty} x e^{-10^{-4}x}\mathrm{d}x \\ &= \frac{1}{10^8}\left| \frac{x \cdot e^{-10^{-4}x}}{-10^{-4}} - \frac{1 \cdot e^{-10^{-4}x}}{(-10^{-4})^2} \right|_{10^4}^{\infty} \\ &= \frac{1}{10^8}\left(\frac{e^{-1}}{10^{-8}} + \frac{e^{-1}}{10^{-8}}\right) \\ &= e^{-1} + e^{-1} \\ &= 2e^{-1} \\ &= 0.7358 \end{aligned}$$

Example 3

In a certain city, the daily consumption of electric power in millions of kilowatt hours can be treated as a random variable having gamma distribution with parameters $\lambda = \frac{1}{2}$ and $r = 3$. If the power plant of this city has a daily capacity of 12 millions kilowatt-hours, what is the probability that this power supply will be inadequate on any given day.

Solution

Let X be a random variable which denotes the daily consumption of electric power in millions killowatt-hours.

$$\begin{aligned} f(x) &= \frac{\lambda^r}{\lceil r} x^{r-1} e^{-\lambda x}, \ x > 0 \\ &= \frac{\left(\frac{1}{2}\right)^3}{\lceil 3} x^2 e^{-\frac{1}{2}x} \end{aligned}$$

$$\begin{aligned} P(\text{power supply is inadequate}) &= P(X > 12) \\ &= \int_{12}^{\infty} f(x)\mathrm{d}x \\ &= \int_{12}^{\infty} \frac{1}{\lceil 3}\frac{1}{2^3} x^2 e^{-\frac{1}{2}x}\mathrm{d}x \\ &= \frac{1}{16}\left| x^2\left(\frac{e^{-\frac{1}{2}x}}{-\frac{1}{2}}\right) - 2x\left(\frac{e^{-\frac{1}{2}x}}{\frac{1}{4}}\right) + 2\left(\frac{e^{-\frac{1}{2}x}}{-\frac{1}{8}}\right) \right|_{12}^{\infty} \\ &= \frac{1}{16} e^{-6}(288 + 96 + 16) \\ &= 25e^{-6} \\ &= 0.062 \end{aligned}$$

Example 4

If a company employs n sales persons, its gross sales in thousands of rupees may be regarded as a random variable having a gamma distribution with $\lambda = \frac{1}{2}$ and $r = 80\sqrt{n}$. If the sales cost is ₹8000 per *salesperson, how many salespersons should the company employ to maximise the expected profit?*

Solution

Let X be the random variable which denotes the gross sales in rupees by n salespersons.

$$\lambda = \frac{1}{2}, \quad r = 80000\sqrt{n}$$

$$f(x) = \frac{\lambda^r}{\Gamma r} x^{r-1} e^{-\lambda x}, \quad x > 0$$

$$E(X) = \frac{r}{\lambda} = \frac{80000\sqrt{n}}{\frac{1}{2}} = 160000\sqrt{n}$$

If y denotes the total expected profit of the company,

$$y = \text{total expected sales} - \text{total sales cost}$$

$$= 160000\sqrt{n} - 8000n$$

$$\frac{dy}{dn} = \frac{80000}{\sqrt{n}} - 8000$$

For maximum profits,

$$\frac{dy}{dn} = 0$$

$$\frac{80000}{\sqrt{n}} - 8000 = 0$$

$$\frac{80000}{\sqrt{n}} = 8000$$

$$\sqrt{n} = 10$$

$$n = 100$$

$$\frac{d^2y}{dn^2} = -\frac{40000}{n^{\frac{3}{2}}}$$

When $n = 100$, $\frac{d^2y}{dn^2} = -40 < 0$

$\therefore$ y is maximum when $n = 100$.

Hence, the company should employ 100 salespersons to maximise the expected profit.

Example 5

Consumer demand for milk in a certain locality, per month, is known to be a general gamma random variable. If the average demand is 'a' litres and the most likely demand is 'b' litres ($b < 0$), what is the variance of the demand?

Solution

Let X be the random variable which denotes the monthly consumer demand of milk. Average demand is the value of $E(X)$. Most likely demand is the value of the mode of X or the value of X for which its probability density function is maximum.

$$f(x) = \frac{\lambda^r}{\Gamma r} x^{r-1} e^{-\lambda x}, \quad x > 0$$

$$f'(x) = \frac{\lambda^r}{\Gamma r}\left[(r-1)x^{r-2}e^{-\lambda x} - \lambda x^{r-1}e^{-\lambda x}\right]$$

$$= \frac{\lambda^r}{\Gamma r} x^{r-2} e^{-\lambda x}\left[(r-1) - \lambda x\right]$$

For maximum value of $f(x)$,

$$f'(x) = 0$$

$$(r-1) - \lambda x = 0$$

$$x = \frac{r-1}{\lambda}$$

Differentiating $f'(x)$ w.r.t. x,

$$f''(x) = \frac{\lambda^r}{\Gamma r}\left[-\lambda x^{r-2}e^{-\lambda x} + \{(r-1) - \lambda x\}\frac{d}{dx}\{x^{r-2}e^{-\lambda x}\}\right]$$

$$= \frac{\lambda^r}{\Gamma r} x^{r-3} e^{-\lambda x}(1-r)$$

$f''(x) < 0$ when $x = \frac{r-1}{\lambda}$

$f(x)$ is maximum when $x = \frac{r-1}{\lambda}$ if $f''(x) < 0$

$$f''(x) < 0 \text{ if } 1 - r < 0$$

$$1 < r$$

or $$r > 1$$

Most likely demand = $\frac{r-1}{\lambda} = b, r > 1$

$$\frac{r-1}{\lambda} = b$$

$$\frac{r}{\lambda} = b + \frac{1}{\lambda} \qquad \text{...(1)}$$

Average demand = $E(X) = \frac{r}{\lambda} = a$ ...(2)

Putting in Eq. (1),

$$a = b + \frac{1}{\lambda}$$

$$\frac{1}{\lambda} = a - b \qquad \text{...(3)}$$

$$\text{Var}(X) = \frac{r}{\lambda^2} = \frac{r}{\lambda} \cdot \frac{1}{\lambda}$$

$$= a(a-b) \qquad \text{[from Eq. (2) and (3)]}$$

EXERCISE 5.5

1. Find the probabilities that the value of a random variable will exceed 4, if it has gamma distribution with

   (a) $\lambda = \frac{1}{3}, r = 2$ (b) $\lambda = \frac{1}{4}, r = 3$

   [Ans.: (a) 0.5551 (b) 4]

2. If X follows the gamma distribution with parameter λ and r, prove that the expected value of the positive square root of X is $\dfrac{\left|\overline{r+\frac{1}{2}}\right.}{\sqrt{\lambda}\left|\overline{r}\right.}$.

3. A random sample of size n is taken from a population which is exponentially distributed with parameter λ. If $\overline{X}$ is the sample mean, show that $n\lambda\overline{X}$ follows a simple gamma distribution with parameter n.

CHAPTER 6

Applied Statistics: Test of Hypothesis

Chapter Outline

6.1 Introduction
6.2 Terms Related to Tests of Hypothesis
6.3 Procedure for Testing of Hypothesis
6.4 Test of Significance for Large Samples
6.5 Test of Significance for Single Proportion – Large Samples
6.6 Test of Significance for Difference between Two Proportions – Large Samples
6.7 Test of Significance for Single Mean – Large Samples
6.8 Test of Significance for Difference between Two Means – Large Samples
6.9 Test of Significance for Difference of Standard Deviations – Large Samples
6.10 Small Sample Tests
6.11 Student's *t*-distribution
6.12 *t*-test: Test of Significance for Single Mean
6.13 *t*-test: Test of Significance for Difference of Means
6.14 *t*-test: Test of Significance for Correlation Coefficients
6.15 Snedecor's *F*-test for Ratio of Variances
6.16 Chi-square (χ^2) Test
6.17 Chi-square Test: Goodness of Fit
6.18 Chi-square Test for Independence of Attributes

6.1 INTRODUCTION

The main purpose behind the sampling theory is the study of the Tests of Hypothesis or Tests of significance. In many situations, assumptions are made about the population parameters involved in order to arrive at decisions related to population on the basis of sample information. Such an assumption is called statistical hypothesis which may or may not be true. The procedure which enables us to decide on the basis of sample results whether a hypothesis is true or not, is called test of hypothesis or test of significance.

6.2 TERMS RELATED TO TESTS OF HYPOTHESIS

(1) *Parameters*: The statistical constants of population such as mean (μ), standard deviation (σ), correlation coefficient (ρ), population proportion (P) etc. are called the parameters. Greek letters are used to denote the population parameters.

(2) *Statistic*: The statistical constants for the sample drawn from the given population such as mean ($\bar{x}$), standard deviation (s), correlation coefficient (r), sample proportion (p) etc., are called the statistic. Roman letters are used to denote the sample statistic.

(3) *Sampling Distribution*: Consider all possible samples of size 'n' which can be drawn from a population of size 'N'. These samples will give different values of a statistic. The means of the samples will not be identical. If these different means are arranged according to their frequencies, the frequency distribution formed is called sampling distribution of mean. Similarly, the sampling distribution of other statistics can be defined.

(4) *Standard Error*: The standard deviation of the sampling distribution of a statistic is known as its standard error SE. Standard error plays a very important role in the large sample theory and forms the basis of the testing of hypothesis.

(5) *Null Hypothesis*: Null hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true. It is denoted by H_0. It asserts that there is no significant difference between the statistic and the population parameter and whatever observed difference exists, is merely due to the fluctuations in sampling from the same population.

(6) *Alternative Hypothesis*: Any hypothesis which is complementary to the null hypothesis is called an alternative hypothesis. It is denoted by H_1. It is set in such a way that the rejection of null hypothesis implies the acceptance of alternative hypothesis. For example, if the null hypothesis is that the average height of the students of a college is 166 cm. i.e., $\mu_0 = 166$ cm, say then the null hypothesis is

$$H_0 : \mu = 166 (= \mu_0)$$

and the alternative hypothesis could be

(i) $H_1 : \mu \neq \mu_0$ (i.e., $\mu > \mu_0$ or $\mu < \mu_0$)

(ii) $H_1 : \mu > \mu_0$

(iii) $H_1 : \mu < \mu_0$

Thus, there can be more than one alternative hypothesis.

(7) *Test Statistic*: After setting up the null hypothesis and alternative hypothesis, test statistic is calculated. The test statistic is a statistic based on appropriate probability distribution. It is used to test whether the null hypothesis should be accepted or rejected. Different probability distribution values are used in appropriate cases while testing the null hypothesis.
For Z-distribution under normal curve for large samples ($n > 30$), the Z-statistic is defined by

$$Z = \frac{t - E(t)}{SE(t)}$$

(8) *Errors in Hypothesis Testing*: The main objective in sampling theory is to draw valid inferences about the population parameters on the basis of the sample results. There is every chance that a decision regarding a null hypothesis may be correct or may not be correct. There are two types of errors.

(i) *Type I error*: It is the error of rejecting the null hypothesis H_0, when it is true. It occurs when a null hypothesis is true, but the difference of means is significant and the hypothesis is rejected. If the probability of making a type I error is denoted by α, the level of significance, then the probability of making a correct decision is $(1 - \alpha)$.

(ii) *Type II error*: It is the error of accepting the null hypothesis H_0, when it is false. It occurs when a null hypothesis is false, but the difference of means is insignificant and the hypothesis is accepted. The probability of making a type II error is denoted by β.

(9) *Level of Significance*: The level of significance is the maximum probability of making a type I error and is denoted by α, i.e., P(Rejecting H_0 when H_0 is true) $= \alpha$. The commonly used level of significance in practice are 5% (0.05) and 1% (0.01). For 5% level of significance ($\alpha = 0.05$), the probability of making type I error is 0.05 or 5% i.e., P(Rejecting H_0 when H_0 is true) = 0.05. This means that there is a probability of making 5 out of 100 type I error. Similarly, 1% level of significance ($\alpha = 0.01$) means that there is a probability of making 1 error out of 100. If no level of significance is given, α is taken as 0.05.

(10) *Critical Region*: The critical region or rejection region is the region of the standard normal curve corresponding to a predetermined level of significance α. The region under the normal curve which is not covered by the rejection region is known as acceptance region. Thus, the statistic which leads to rejection of null hypothesis H_0 gives rejection region or critical region. The value of the test statistic calculated to test the null hypothesis H_0 is known as critical value. Thus, the critical value separates the rejection region from the acceptance region.

(11) *Two Tailed Test and One Tailed Test:* When the test of hypothesis is made on the basis of rejection region represented by both the sides of the standard normal curve, it is called a two tailed test. A test of statistical hypothesis, where the alternative hypothesis H_1 is two sided or two tailed such as:

Null Hypothesis $H_0 : \mu = \mu_0$

Alternative Hypothesis $H_1 : \mu \neq \mu_0$ ($\mu > \mu_0$ and $\mu < \mu_0$), is called two tailed test or two sided test.

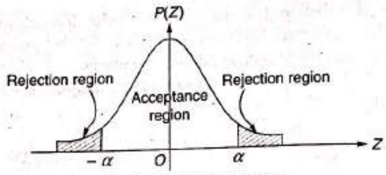


Fig. 6.1 Two tailed test

A test of statistical hypothesis, where the alternative hypothesis is one sided is called one tailed test or one sided test. There are two types of one tailed tests.

(i) *Right Tailed Test:* In the right tailed test, the rejection region or critical region lies entirely on the right tail of the normal curve (Fig. 6.2).

(ii) *Left Tailed Test:* In the left tailed test, the rejection region or critical region lies entirely on the left tail of the normal curve.

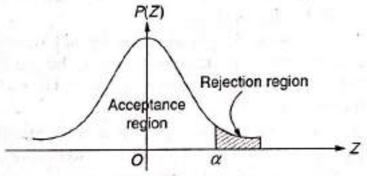


Fig. 6.2 Right tailed test

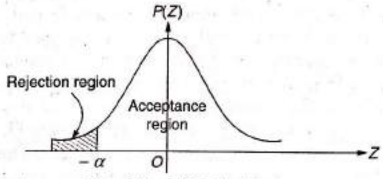


Fig. 6.3 Left tailed test

For example, in a test for testing the mean (μ) of the population Null Hypothesis $H_0 : \mu = \mu_0$

Alternative Hypothesis $H_1 : \mu > \mu_0$ (Right tailed)

$\mu < \mu_0$ (Left tailed)

A two tailed test is applied in such cases when the difference between the sample mean and population mean is tending to reject the null hypothesis H_0, the difference may be positive or negative.

A one tailed test is applied in such cases when the population mean is at least as large as some specified value of the mean (right tailed test) or at least as small as some specified value of the mean (left tailed test).

Critical value (Z_α)	*Level of significance* α		
	1%	5%	10%
Two tailed test	$\lvert Z_\alpha \rvert = 2.58$	$\lvert Z_\alpha \rvert = 1.96$	$\lvert Z_\alpha \rvert = 1.645$
Right tailed test	$Z_\alpha = 2.33$	$Z_\alpha = 1.645$	$Z_\alpha = 1.28$
Left tailed test	$Z_\alpha = -2.33$	$Z_\alpha = -1.645$	$Z_\alpha = -1.28$

(12) *Confidence Limits:* The limits within which a hypothesis should lie with specified probability are called confidence limits or fiducial limits. Generally, the confidence limits are set up with 5% or 1% level of significance. If the sample value lies between the confidence limits, the hypothesis is accepted, if it does not, then the hypothesis is rejected at the specified level of significance. Suppose that the sampling distribution of a statistic S is normal with mean μ and standard deviation σ. The sample statistic S can be expected to lie in the interval ($\mu - 1.96\sigma$, $\mu + 1.96\sigma$) for 95% times (Fig. 6.29). Because of this, ($S - 1.96\sigma$, $S + 1.96\sigma$) is called the 95% confidence interval for estimation of μ. The ends of this interval, i.e., $S \pm 1.96\sigma$ are called 95% confidence limits for S. Similarly, $S \pm 2.58\sigma$ are 99% confidence limits. The numbers 1.96, 2.58 etc. are called confidence coefficients.

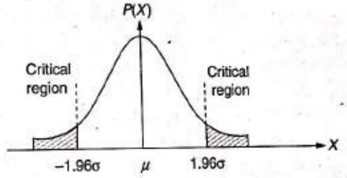


Fig. 6.4 Confidence Limits

6.3 PROCEDURE FOR TESTING OF HYPOTHESIS

The various steps in testing of a statistical hypothesis are as follows:

(i) *Null Hypothesis*: Set up the Null Hypothesis H_0

(ii) *Alternative Hypothesis*: Set up the Alternative Hypothesis H_1. This will decide the use of single-tailed (right or left) or Two tailed test.

(iii) *Level of Significance*; Select the appropriate level of significance (α) depending on the reliability of the estimates and permissible risk. If no level of significance is given, α is selected as 0.05.

(iv) *Test Statistic*: Calculate the test statistic

$$Z = \frac{t - E(t)}{SE(t)} \text{ under } H_0$$

(v) *Critical Value*: Find the significant value (tabulated value) Z_α of Z at the given level of significance α.

(vi) *Decision*: Compare the calculated value of Z with the tabulated value Z_α. If $|Z| < Z_\alpha$ i.e., if the calculated value of Z is less than tabulated value Z_α at the level of significance α, the null hypothesis is accepted. If $|Z| > Z_\alpha$ i.e., if the calculated value of Z is more than tabulated value Z_α at the level of significance α, the null hypothesis is rejected.

6.4 TEST OF SIGNIFICANCE FOR LARGE SAMPLES

If a sample consists of more than 30 items, i.e., $n > 30$, it is considered as large sample. The following assumptions are applied for significance tests of large samples:

(i) The random sampling distribution of statistic has the properties of the normal curve.

(ii) Values (i.e., statistic) given by the samples are sufficiently close to the populate values (i.e., parameters) and can be used in its place for calculating the standard error (*SE*) of the estimate.

For example, if SD of the population is not known, SE can be calculated by SD of the sample.

Suppose the hypothesis to be tested is that the probability of success in such trail is p. Assuming it to be true, the mean μ and the standard deviation σ of the sampling distribution of the number of successes are np and $\sqrt{npq}$ respectively as the sampling distribution of number of successes follows a binomial probability distribution.

If x is the observed number of successes in the sample and Z is the standard normal variate then

$$Z = \frac{x-\mu}{\sigma}$$

The tests of significance are as follows:

(i) If $|Z| < 1.96$, the difference between the observed and expected number of successes is not significant.

(ii) If $|Z| > 1.96$, the difference is significant at 5% level of significance.

(iii) If $|Z| > 2.58$, the difference is significant at 1% level of significance.

Example 1

A coin was tossed 960 *times and returned heads* 183 *times. Test the hypothesis that the coin is unbiased. Use a* 0.05 *level of significance.*

Solution

$$n = 960$$

$$p = \text{probability of getting head} = \frac{1}{2}$$

$$q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2}$$

$$\mu = np = 960\left(\frac{1}{2}\right) = 480$$

$$\sigma = \sqrt{npq} = \sqrt{960 \times \frac{1}{2} \times \frac{1}{2}} = 15.49$$

$$x = \text{number of sucesses} = 183$$

(i) Null Hypothesis H_0: The coin is unbiased.

(ii) Alternative Hypothesis H_1: The coin is biased.

(iii) Level of significance: $\alpha = 0.05$

(iv) Test statistic: $Z = \frac{x-\mu}{\sigma} = \frac{183-480}{15.49} = -19.17$

$$|Z| = 19.17$$

(v) Critical value: $|Z_{0.05}| = 1.96$

(vi) Decision: Since $|Z| > |Z_{0.05}|$, the null hypothesis is rejected at 5%-level of significance, i.e., the coin is biased.

Example 2

A dice is tossed 960 *times and it falls with* 5 *upwards* 184 *times. Is the dice unbiased at a level of significance of* 0.01*?*

Solution

$$n = 960$$

$$p = \text{Probability of throwing 5 with one die} = \frac{1}{6}$$

$$q = 1 - p = 1 - \frac{1}{6} = \frac{5}{6}$$

$$\mu = np = 960\left(\frac{1}{6}\right) = 160$$

$$\sigma = \sqrt{npq} = \sqrt{960 \times \frac{1}{6} \times \frac{5}{6}} = 11.55$$

$$x = \text{number of successes} = 184$$

(i) Null Hypothesis H_0: The dice is unbiased.

(ii) Alternative Hypothesis H_1: The dice is biased.

(iii) Level of significance: $\alpha = 0.01$

(iv) Test statistic: $Z = \frac{x-\mu}{\mu} = \frac{184-160}{11.55} = 2.08$

$$|Z| = 2.08$$

(v) Critical value: $|Z_{0.01}| = 2.58$

(vi) Decision: Since $|Z| < |Z_{0.01}|$, the null hypothesis is accepted at 1% level of significance, i.e., the dice is unbiased.

6.5 TEST OF SIGNIFICANCE FOR SINGLE PROPORTION – LARGE SAMPLES

Let p be the sample proportion in a large random sample of size n drawn from a population having proportion P. Also, the population proportion P has a specified value P_0.

Working Rule

(i) Null Hypothesis H_0: $P = P_0$, i.e., the population proportion P has a specified value P_0.

(ii) Alternative Hypothesis H_1: $P \neq P_0$ (i.e., $P > P_0$ or $P < P_0$)
or H_1: $P > P_0$
or H_1: $P < P_0$

(iii) Level of significance: Select the level of significance α

(iv) Test statistic: $Z = \dfrac{p-P}{\sqrt{\dfrac{PQ}{n}}}$, where $Q = 1 - P$

(v) Critical Value: Find the critical value (tabulated value) Z_α of Z at the given level of significance.

(vi) Decision: If $|Z| < Z_\alpha$ at the level of significance α, the null hypothesis is accepted. If $|Z| > Z_\alpha$ at the level of significance α, the null hypothesis is rejected.

Note

1. Null Hypothesis H_0 is rejected when $|Z| > 3$ without mentioning any level of significance.
2. Confidence limits:
   (i) 95% confidence limits = $p \pm 1.96\sqrt{\dfrac{PQ}{n}}$
   (ii) 99% confidence limits = $p \pm 2.58\sqrt{\dfrac{PQ}{n}}$

If the population proportions P and Q are not known, p and q are used in equations.

Example 1

A manufacturer claimed that atleast 95% of the equipment which he supplied to a factory conformed to specification. An examination of a sample of 200 pieces of equipment revealed that 18 were faulty. Test his claim at 5% level of significance.

Solution

$n = 200$

Number of pieces conforming to specification = 200 − 18 = 182

p = Sample proportion of pieces conforming to specification = $\dfrac{182}{200} = 0.91$

P = Population proportion of pieces conforming to specification = 0.95
$Q = 1 - P = 1 - 0.95 = 0.05$

(i) *Null Hypothesis* H_0: $P = 0.95$ i.e., the proportion of pieces conforming to proportion is 95%.

(ii) *Alternating Hypothesis* H_1: $P < 0.95$ (Left tailed test)

(iii) *Level of significance*: $\alpha = 0.05$

(iv) Test statistic: $Z = \dfrac{p-P}{\sqrt{\dfrac{PQ}{n}}} = \dfrac{0.91-0.95}{\sqrt{\dfrac{(0.95)(0.05)}{200}}} = -2.59$

$$|Z| = 2.59$$

(v) Critical value: $|Z_{0.05}| = 1.645$

(vi) Decision: Since $|Z| > |Z_{0.05}|$, the null hypothesis is rejected at 5% level of significance, i.e., the manufacturer's claim is rejected.

Example 2

In a hospital 480 *female and* 520 *male babies were born in a week. Do these figures confirm the hypothesis that males and females were born in equal numbers?*

Solution

n = Total number of births = 480 + 520 = 1000

p = Sample proportion of females born = $\dfrac{480}{1000} = 0.48$

P = Population proportion of females born = 0.5
$Q = 1 - P = 1 - 0.5 = 0.5$

(i) Null Hypothesis H_0: $P = 0.5$ i.e., the males and females were born in equal numbers.

(ii) Alternative Hypothesis H_1: $P \neq 0.5$ (Two tailed test)

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: $Z = \dfrac{p-P}{\sqrt{\dfrac{PQ}{n}}} = \dfrac{0.48-0.5}{\sqrt{\dfrac{(0.5)(0.5)}{1000}}} = -1.265$

$$|Z| = 1.265$$

(v) Critical value: $|Z_{0.05}| = 1.96$

(vi) Decision: Since $|Z| < |Z_{0.05}|$, the null hypothesis is accepted at 5% level of significance, i.e., males and females were born in equal proportions.

Example 3

In a study designed to investigate whether certain detonators used with explosives in a coal mining meet the requirement that at least 90% *will ignite the explosive when charged. It is found that* 174 *of* 200 *detonators function properly. Test the null hypothesis* $P = 0.9$ *against the alternative hypothesis* $P < 0.9$ *at the* 0.05 *level of significance.*

Solution

$$n = 2000$$

p = Sample proportion of detonators functioning properly $= \frac{174}{200} = 0.87$

P = Population proportion of detonators functioning properly = 0.9

$$Q = 1 - P = 1 - 0.9 = 0.1$$

(i) Null Hypothesis H_0: $P = 0.9$

(ii) Alternative Hypothesis H_1: $P < 0.9$ (Left tailed test)

(iii) Level of significance: $\alpha = 0.05$

(iv) Test statistic: $Z = \frac{p-P}{\sqrt{\frac{PQ}{n}}} = \frac{0.87 - 0.9}{\sqrt{\frac{(0.9)(0.1)}{200}}} = -1.41$

$$|Z| = 1.41$$

(v) Critical value: $|Z_{0.05}| = 1.645$

(vi) Decision: Since $|Z| < |Z_{0.05}|$, the null hypothesis is accepted at 5% level of significance.

Example 4

A salesman in a departmental store claims that at most 60 *percent of the shoppers entering the store leave without making a purchase. A random sample of* 50 *shoppers showed that* 35 *of them left without making a purchase. Are these sample results consistent with the claim of the salesman? Use a level of significance of* 0.05.

Solution

$$n = 50$$

p = Sample proportion of shoppers not making a purchase $= \frac{35}{50} = 0.7$

p = Population proportion of shoppers not making a purchase = 0.6

$$Q = 1 - P = 1 - 0.6 = 0.4$$

(i) Null Hypothesis H_0: $P = 0.6$, i.e., the proportion of shoppers not making a purchase is 60%.

(ii) Alternative Hypothesis H_1: $P > 0.6$ (Right tailed test)

(iii) Level of significance: $\alpha = 0.05$

(iv) Test statistic: $Z = \frac{p-P}{\sqrt{\frac{PQ}{n}}} = \frac{0.7 - 0.6}{\sqrt{\frac{(0.6)(0.4)}{50}}} = 1.443$

$$|Z| = 1.443$$

(v) Critical value: $Z_{0.05} = 1.645$

(vi) Decision: Since $|Z| < Z_{0.05}$, the null hypothesis is accepted, i.e., the sample results are consistent with claim of the salesman.

Example 5

The fatality rate of typhoid patients is believed to be 17.26%. *In a certain year* 640 *patients suffering from typhoid were treated in a metropolitan hospital and only* 63 *patients died. Can you consider the hospital efficient at* 1% *level of significance?*

Solution

$$n = 640$$

p = Sample proportion of typhoid patients died $= \frac{63}{640} = 0.0984$

P = Population proportion of typhoid patients died = 0.1726

$$Q = 1 - P = 1 - 0.1726 = 0.8274$$

(i) Null Hypothesis H_0: $P = 0.1726$, i.e., the hospital is efficient.

(ii) Alternative Hypothesis H_1: $P < 0.1726$ (Left tailed test)

(iii) Level of significance: $\alpha = 0.01$

(iv) Test statistic: $Z = \frac{p-P}{\sqrt{\frac{PQ}{n}}} = \frac{0.0984 - 0.1726}{\sqrt{\frac{(0.1726)(0.8274)}{640}}} = -4.97$

$$|Z| = 4.97$$

(v) Critical value: $|Z_{0.01}| = 2.33$

(vi) Decision: Since $|Z| > |Z_{0.01}|$, the null hypothesis is rejected at 1% level of significance, i.e., the hospital is efficient.

Example 6

In a big city, 325 men out of 600 were found to be smokers. Does this information support the conclusion that the majority of men in this city are smokers?

Solution

$$n = 600$$

$$p = \text{Sample proportion of smokers in city} = \frac{325}{600} = 0.542$$

$$P = \text{Population proportion of smokers in city} = 0.5$$

$$Q = 1 - P = 1 - 0.5 = 0.5$$

(i) Null Hypothesis H_0: $P = 0.5$, i.e., the proportion of smokers in the city is 50%.

(ii) Alternative Hypothesis H_1: $P > 0.5$ (Right tailed test)

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: $Z = \dfrac{p - P}{\sqrt{\dfrac{PQ}{n}}} = \dfrac{0.542 - 0.5}{\sqrt{\dfrac{(0.5)(0.5)}{600}}} = 2.06$

$$|Z| = 2.06$$

(v) Critical value: $Z_{0.05} = 1.645$

(vi) Decision: Since $|Z| > Z_{0.05}$, the null hypothesis is rejected at 5% level of significance, i.e., proportion of smokers in city is more than 50% and majority of men in the city are smokers.

Example 7

In a random sample of 160 worker exposed to a certain amount of radiation, 24 experienced some ill effects. Construct a 95% confidence interval for the corresponding true percentage.

Solution

$$n = 160$$

$$p = \text{Sample proportion of workers exposed to radiation} = \frac{24}{160} = 0.15$$

$$q = 1 - p = 1 - 0.15 = 0.85$$

Confidence interval at 95% level of significance is

$$\left(p - 1.96\frac{\sqrt{pq}}{n}, p + 1.96\sqrt{\frac{pq}{n}}\right)$$

i.e., $$\left(0.15 - 1.96\sqrt{\frac{(0.15)(0.85)}{160}}, 0.15 + 1.96\sqrt{\frac{(0.15)(0.85)}{160}}\right)$$

i.e., $$(0.0947, 0.2053)$$

6.6 TEST OF SIGNIFICANCE FOR DIFFERENCE OF PROPORTIONS – LARGE SAMPLES

Let p_1 and p_2 be the sample proportions in two large samples of sizes n_1 and n_2 drawn from two populations having proportions P_1 and P_2.

Working Rule

(i) Null Hypothesis H_0: $P_1 = P_2$, i.e., there is no significant difference in two population proportions P_1 and P_2.

(ii) Alternative Hypothesis H_1: $P_1 \neq P_2$
or H_1: $P_1 > P_2$
or H_1: $P_1 < P_2$

(iii) Level of significance: Select level of significance α

(iv) Test statistic: There are two cases:

(a) When the population proportions P_1 and P_2 are known

$$Z = \frac{P_1 - P_2}{\sqrt{\dfrac{P_1Q_1}{n_1} + \dfrac{P_2Q_2}{n_2}}}$$

(b) When the population proportions P_1 and P_2 are not known but sample proportions p_1 and p_2 are known

There are two methods to estimate P_1 and P_2.
Method of Substitution: In this method, sample proportions p_1 and p_2 are substituted for P_1 and P_2.

$$Z = \frac{p_1 - p_2}{\sqrt{\dfrac{p_1q_1}{n_1} + \dfrac{p_2q_2}{n_2}}}$$

Method of pooling: In this method, the estimated value of two population proportions is obtained by pooling the two sample proportions p_1 and p_2 into a single proportion p.

$$p = \frac{n_1 p + n_2 p_2}{n_1 + n_2}$$

$$Z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

(v) Critical value: Find the critical value (tabulated value) of Z at given level of significance.

(vi) Decision: If $|Z| < Z_\alpha$ at the level of significance, the null hypothesis is accepted. If $|Z| > Z_\alpha$ at the level of significance, the null hypothesis is rejected.

Note

1. Null Hypothesis H_0 is rejected when $|Z| > 3$ without mentioning any level of significance.
2. Confidence limits:
   (i) 95% confidence limits $= (p_1 - p_2) \pm 1.96\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$
   (ii) 99% confidence limits $= (p_1 - p_2) \pm 2.58\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$

If the population proportions P_1 and P_2 are not known, p_1, p_2, q_1 and q_2 are used in equations.

Example 1

Random samples of 400 men and 600 women were asked whether they would like to have a flyover near their residence 200 men and 325 women were in favour of the proposal. Test the hypothesis that proportions of men and women in favour of the proposal are same at 5% level of significance.

Solution

$$n_1 = 400,\ n_2 = 600$$

$$p_1 = \text{Proportion of men} = \frac{200}{400} = 0.5$$

$$p_2 = \text{Proportion of women} = \frac{325}{600} = 0.541$$

$$p = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{(400)(0.5) + (600)(0.541)}{400 + 600} = 0.525$$

$$q = 1 - p = 1 - 0.525 = 0.475$$

(i) Null Hypothesis H_0: $P_1 = P_2$, i.e., there is no significant difference in proportion of men and women in favour of the proposal.

(ii) Alternative Hypothesis is H_1: $P_1 \neq P_2$ (Two tailed test)

(iii) Level of significance: $\alpha = 0.05$

(iv) Test statistic: $Z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.5 - 0.541}{\sqrt{(0.525)(0.475)\left(\frac{1}{400} + \frac{1}{600}\right)}} = -1.28$

$$|Z| = 1.28$$

(v) Critical value: $|Z_{0.05}| = 1.96$

(vi) Decision: Since $|Z| < |Z_{0.05}|$, the null hypothesis is accepted at 5% level of significance, i.e., there is no significant difference of opinion between men and women in favour of the proposal.

Example 2

In a city A, 20% of a random sample of 900 school boys has a certain slight physical defect. In another city B, 18.5% of a random sample of 1600 school boys has the same defect. Is the difference between the proportions significant at 0.05 level of significance?

Solution

$$n_1 = 900,\ n_2 = 1600$$

p_1 = Proportion of school boys in city $A = 0.2$

p_2 = Proportion of school boys in city $B = 0.185$

$$p = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{(900)(0.2) + (1600)(0.185)}{900 + 1600} = 0.1904$$

$$q = 1 - p = 1 - 0.1904 = 0.8096$$

(i) Null Hypothesis H_0: $P_1 = P_2$, i.e., there is no significant difference in proportion of two city school boys.

(ii) Alternative Hypothesis H_1: $P_1 \neq P_2$ (Two tailed test)

(iii) Level of significance: $\alpha = 0.05$

(iv) Test statistic: $Z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.2 - 0.185}{\sqrt{(0.1904)(0.8096)\left(\frac{1}{900} + \frac{1}{1600}\right)}} = 0.916$

$$|Z| = 0.916$$

(v) Critical value: $|Z_{0.05}| = 1.96$

(vi) Decision: Since $|Z| < |Z_{0.05}|$, the null hypothesis is accepted at 5% level of significance, i.e., there is no significant difference between the proportions of two city school boys.

Example 3

Before an increase in excise duty on tea, 800 people out of a sample of 1000 were consumers of tea. After an increase in excise duty, 800 people were consumers of tea in a sample of 1200 persons. Find whether there is significant decrease in the consumption of tea after the increase in duty.

Solution

$$n_1 = 1000,\ n_2 = 1200$$

p_1 = Proportion of consumers of tea before increase in excise duty $= \dfrac{800}{1000} = 0.8$

p_2 = Proportion of consumers of tea after increase in excise duty $= \dfrac{800}{1200} = 0.67$

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{(1000)(0.8) + (1200)(0.67)}{1000 + 1200} = 0.73$$

$$q = 1 - p = 1 - 0.73 = 0.27$$

(i) Null Hypothesis H_0: $P_1 = P_2$, i.e., there is no significant decrease in the consumption of tea after the increase in duty.

(ii) Alternative Hypothesis H_1: $P_1 > P_2$ (Right tailed test)

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: $Z = \dfrac{p_1 - p_2}{\sqrt{pq\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} = \dfrac{0.8 - 0.67}{\sqrt{(0.73)(0.27)\left(\dfrac{1}{1000} + \dfrac{1}{1200}\right)}} = 6.84$

$$|Z| = 6.84$$

(v) Critical value: $Z_{0.05} = 1.645$

(vi) Decision: Since $|Z| > Z_{0.05}$, the null hypothesis is rejected at 5% level of significance, i.e., there is significant decrease in the consumption of tea after the increase in duty.

Example 4

15.5% of a random sample of 1600 undergraduates smokers, whereas 20% of a random sample of 900 postgraduates were smokers in a state. Can we conclude that less number of undergraduates are smokers than the postgraduates?

Solution

$$n_1 = 1600,\ n_2 = 900$$

p_1 = Proportion of undergraduate smokers = 0.155

p_2 = Proportion of postgraduate smokers = 0.2

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{(1600)(0.155) + (900)(0.2)}{1600 + 900} = 0.1712$$

$$q = 1 - p = 1 - 0.1712 = 0.8288$$

(i) Null Hypothesis H_0: $P_1 = P_2$, i.e., there is no significant difference in proportion of undergraduate and postgraduate smokers.

(ii) Alternative Hypothesis H_1: $P_1 < P_2$ (Left tailed test)

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: $Z = \dfrac{p_1 - p_2}{\sqrt{pq\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} = \dfrac{0.155 - 0.2}{\sqrt{(0.1712)(0.8288)\left(\dfrac{1}{1600} + \dfrac{1}{900}\right)}} = -2.87$

$$|Z| = 2.87$$

(v) Critical value: $|Z_{0.05}| = 1.645$

(vi) Decision: Since $|Z| > |Z_{0.05}|$, the null hypothesis is rejected at 5% level of significance, i.e., less number of undergraduates smokers than the postgraduates.

Example 5

A machine produced 20 defective articles in a batch of 400. After overhauling it produced 10 defective articles in a batch of 300. Has the machine improved?

Solution

$$n_1 = 400,\ n_2 = 300$$

p_1 = Proportion of defective articles before overhauling $= \dfrac{20}{400} = 0.05$

p_2 = Proportion of defective articles after overhauling $= \dfrac{10}{300} = 0.033$

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{(400)(0.05) + (300)(0.033)}{400 + 300} = 0.043$$

$$q = 1 - p = 1 - 0.043 = 0.957$$

(i) Null Hypothesis H_0: $P_1 = P_2$, i.e., the proportions of defective articles before and after overhauling are equal.

(ii) Alternative Hypothesis H_1: $P_1 > P_2$ (Right tailed test)

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: $Z = \dfrac{p_1 - p_2}{\sqrt{pq\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}} = \dfrac{0.05-0.033}{\sqrt{(0.043)(0.957)\left(\dfrac{1}{400}+\dfrac{1}{300}\right)}} = 1.097$

$|Z| = 1.097$

(v) Critical value: $Z_{0.05} = 1.645$

(vi) Decision: Since $|Z| < Z_{0.05}$, the null hypothesis is accepted at 5% level of significance, i.e., proportion of defective articles before and after are equal and machine has not improved.

Example 6

In two large populations, there are 30% and 25% fair haired people respectively. Is this difference likely to be hidden in samples of 1200 and 900 respectively from the two populations?

Solution

$n_1 = 1200, n_2 = 900$

P_1 = Proportion of faired people in the first population = 0.3

$Q_1 = 1 - P_1 = 1 - 0.3 = 0.7$

P_2 = Proportion of faired people in the second population = 0.25

$Q_2 = 1 - P_2 = 1 - 0.25 = 0.75$

(i) Null Hypothesis H_0: $P_1 = P_2$, i.e., the difference in population proportions is likely to be hidden in sampling.

(ii) Alternative Hypothesis H_1: $P_1 \neq P_2$ (Two tailed test)

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: $Z = \dfrac{P_1 - P_2}{\sqrt{\dfrac{P_1Q_1}{n_1}+\dfrac{P_2Q_2}{n_2}}} = \dfrac{0.3-0.25}{\sqrt{\dfrac{(0.3)(0.7)}{1200}+\dfrac{(0.25)(0.75)}{900}}} = 2.56$

$|Z| = 2.56$

(v) Critical value: $|Z_{0.05}| = 1.96$

(vi) Decision: Since $|Z| > |Z_{0.05}|$, the null hypothesis is rejected at 5% level of significance, i.e., the difference in population proportions is not likely to be hidden in sampling.

Example 7

A random sample of 300 shoppers at a supermarket includes 204 who regularly uses cents off coupons. In another sample of 500 shoppers at a supermarket includes 75 who regularly uses cents off coupons. Obtain 95% confidence limits for the difference in the population proportions.

Solution

$$n_1 = 300,\ n_2 = 500$$

p_1 = Proportion of shoppers who uses cents of coupons in the first sample

$= \dfrac{204}{300} = 0.68$

$q_1 = 1 - p_1 = 1 - 0.68 = 0.32$

p_2 = Proportion of shoppers who uses cents of coupons in the second sample

$= \dfrac{75}{500} = 0.15$

$q_2 = 1 - p_2 = 1 - 0.15 = 0.85$

$$SE = \sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}} = \sqrt{\frac{(0.68)(0.32)}{300}+\frac{(0.15)(0.85)}{500}} = 0.031$$

95% confidence limits for the difference in population proportion is

$$(p_1 - p_2) - 1.96\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}},\ (p_1 - p_2) + 1.96\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}$$

i.e., $(0.68 - 0.15) - 1.96\,(0.031), (0.68 - 0.15) + 1.96\,(0.031)$

i.e., $(0.469, 0.591)$

EXERCISE 6.1

1. A manufacturer claims at least 95% of the items he produces are failure free. Examinations of a random sample of 600 items showed 39 to be defective. Test the claim at a significance level of 0.05.
   [Ans.: Claim is rejected]
2. In a sample of 400 parts manufactured by a factory, the number of defective parts was found to be 30. The company, however, claim that only 5% of their product is defective. Is the claim tenable?
   [Ans.: Claim is rejected]
3. A sample of 600 persons selected at random from a large city shows that the percentage of male in the sample is 53%. It is believed that the male to the total population ratio in the city is $\frac{1}{2}$. Test whether this

belief is confirmed by the observation.

[**Ans.:** Belief is confirmed by the observation]

4. In a sample of 1000 people in Karnataka, 540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular in this state at 1% level of significance?

   [**Ans.:** Both rice and wheat are equally popular in state]

5. In a big city 325 men out of 600 men were found to be smokers. Does this information support the conclusion that the majority of men in this city are smokers?

   [**Ans.:** Majority of men in the city are smokers]

6. A dice was thrown 400 times and 'six' resulted 80 times. Do the data justify the hypothesis of an unbiased dice.

   [**Ans.:** The dice is unbiased]

7. In a random sample of 125 cold drinkers, 68 said they prefer 'Thumsup' to Pepsi'. Test the null hypothesis $P = 0.5$ against the alternative hypothesis $P > 0.5$.

   [**Ans.:** Null hypothesis is accepted]

8. A social worker believes that fewer than 25% of the couples in a certain area have ever used any form of birth control. A random sample of 120 couples was contacted. Twenty of them said they have used. Test the belief of the social worker at 0.05 level.

   [**Ans.:** Belief of the social worker is true]

9. 20 people were attacked by a disease and only 18 survived. Will you reject the hypothesis that the survival rate is attacked by this disease is 85% in favour of the hypothesis that is more at 5% level?

   [**Ans.:** The hypothesis is accepted]

10. A manufacturer of electronic equipment subjects samples of two completing brands of transistors to an accelerated performance test. If 45 of 180 transistors of the first kind and 34 of 120 transistors of second kind fail the test, what can be conclude at the level of significance $\alpha = 0.05$ about the difference between the corresponding sample proportion?

    [**Ans.:** The difference between the proportions is not significant]

11. On the basis of their total scores, 200 candidates of a civil service examination are divided into two groups, the upper 30% and the remaining 70%. Consider the first question of the examination. Among the first group, 40 had the correct answer, whereas among the second group, 80 had the correct answer. On the basis of these results, can one conclude that the first question is not good at discriminating ability of the type being examined here?

    [**Ans.:** The first question is good enough at discriminating ability of the type being examined]

12. A company wanted to introduce a new plan of work and a survey was conducted for this purpose. Out of sample of 500 workers in one group, 62% favoured the new plan and another group of sample of 400 workers, 41% were against the new plan. Is there any significant difference between the two groups in their attitude towards the new plan at 5% level of significance?

    [**Ans.:** There is no significant difference between the two groups in their attitude towards the new plan]

13. In a random sample of 1000 persons from town *A*, 400 are found to be consumers of wheat. In a sample of 800 from town *B*, 400 are found to be consumers of wheat. Do these data reveal a significant difference between town *A* and town *B*, so far as the proportion of wheat consumers is concerned?

    [**Ans.:** There is significant difference between town A and town B as the proportion of wheat consumers is concerned]

14. 100 articles from a factory are examined and 10 are found to be defective. Out of 500 similar articles from a second factory 15 are found to be defective. Test the significance between the difference of two proportions at 5% level.

    [**Ans.:** There is a significant difference between the two proportions]

6.7 TEST OF SIGNIFICANCE FOR SINGLE MEAN – LARGE SAMPLES

Let a random sample size n ($n > 30$) has the sample mean $\bar{x}$ and population has the mean μ. Also, the population mean μ has a specified value μ_0.

Working Rule

(i) Null Hypothesis $H_0 : \mu = \mu_0$, i.e., the population mean μ has a specified value μ_0.

(ii) Alternative Hypothesis $H_1 : \mu \neq \mu_0$.

(iii) Level of significance: Select the level of significance α.

(iv) Test statistic: There are two cases for calculating a test statistic Z.

   (a) When the standard deviation σ of population is known

$$Z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

   (b) When the standard deviation σ of population is not known

$$Z = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)}$$

where s is the sample SD.

(v) Critical value: Find the critical value (tabulated value) Z_α of Z at the given level of significance α.

(vi) Decision: If $|Z| < Z_\alpha$ at the level of significance α, the null hypothesis is accepted. If $|Z| > Z_\alpha$ at the level of significance α, the null hypothesis is rejected.

Note

1. Null Hypothesis H_0 is rejected when $|Z| > 3$ without mentioning any level of significance.
2. Confidence limits:
   (i) 95% confidence limits = $\bar{x} \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$

   (ii) 99% confidence limits = $\bar{x} \pm 2.58\left(\frac{\sigma}{\sqrt{n}}\right)$

If standard deviation σ of population is not known, s is used in equations.

Example 1

A random sample of 100 *Indians has an average life span of* 71.8 *years with standard deviation of* 8.9 *years. Can it be concluded that the average life span of an Indian is* 70 *years?*

Solution

$$n = 100,\ \bar{x} = 71.8 \text{ years},\ \mu = 70 \text{ years},\ s = 8.9 \text{ years}$$

(i) Null Hypothesis H_0: $\mu = 70$ years i.e., the average life span of an Indian is 70 years.

(ii) Alternative Hypothesis H_1: $\mu \neq 70$ years (Two tailed test)

(iii) Level of Significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: $Z = \dfrac{\bar{x}-\mu}{\left(\frac{s}{\sqrt{n}}\right)} = \dfrac{71.8-70}{\left(\frac{8.9}{\sqrt{100}}\right)} = 2.02$

$$|Z| = 2.02$$

(v) Critical value: $|Z_{0.05}| = 1.96$

(vi) Decision: Since $|Z| > |Z_{0.05}|$, the null hypothesis is rejected at 5% level of significance, i.e., the average life span of an Indian is not 70 years.

Example 2

A random sample of 50 *items gives the mean* 6.2 *and variance* 10.24. *Can it be regarded as drawn from a normal population with mean* 5.4 *at* 5% *level of significance?*

Solution

$$n = 50,\quad \bar{x} = 6.2,\quad \mu = 5.4,\quad s = \sqrt{10.24}$$

(i) Null Hypothesis H_0: $\mu = 5.4$, i.e., the sample is drawn from a normal population with mean 5.4.

(ii) Alternative Hypothesis H_1: $\mu \neq 5.4$ (Two tailed test)

(iii) Level of significance: $\alpha = 0.05$

(iv) Test statistic: $Z = \dfrac{\bar{x}-\mu}{\left(\frac{s}{\sqrt{n}}\right)} = \dfrac{6.2-5.4}{\left(\frac{\sqrt{10.24}}{\sqrt{50}}\right)} = 1.77$

$$|Z| = 1.77$$

(v) Critical value: $|Z_{0.05}| = 1.96$

(vi) Decision: Since $|Z| < |Z_{0.05}|$, the null hypothesis is accepted at 5% level of significance i.e., the sample is drawn from a normal population with mean 5.4.

Example 3

A random sample of 400 *members is found to have a mean of* 4.45 *cm. Can it be reasonably regarded as a sample from a large population whose mean is* 5 *cm and variance is* 4 *cm?*

Solution

$$n = 400,\quad \bar{x} = 4.45\text{cm},\quad \mu = 5\text{cm}\quad \sigma = \sqrt{4} = 2\text{ cm}$$

(i) Null Hypothesis H_0: $\mu = 5$ cm, i.e., the sample is drawn from a large population with mean 5 cm.

(ii) Alternative Hypothesis H_1: $\mu \neq 5$ cm (Two tailed test)

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: $Z = \dfrac{\bar{x}-\mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \dfrac{4.45-5}{\left(\frac{2}{\sqrt{400}}\right)} = 5.55$

$$|Z| = 5.55$$

(v) Critical value: $|Z_{0.05}| = 1.96$

(vi) Decision: Since $|Z| > |Z_{0.05}|$, the null hypothesis is rejected at 5% level of significance, i.e., the sample is not drawn from the large population with mean 5 cm.

Example 4

A sample of 900 *members has a mean of* 3.4 *cm and SD* 2.61 *cm. Is the sample from a large population of mean* 3.25 *cm and SD* 2.61 *cm? If the population is normal and its mean is unknown, find the 95% fiducial limits of its true mean.*

Solution

$$n = 900, \bar{x} = 3.4 \text{ cm}, s = 2.61 \text{ cm}, \mu = 3.25 \text{ cm}, \sigma = 2.61 \text{ cm}$$

(i) Null Hypothesis H_0: $\mu = 3.25$ cm, i.e., the sample has been drawn from the population with mean $\mu = 3.25$ cm and SD = 2.61 cm.

(ii) Alternative Hypothesis H_1: $\mu \neq 3.25$ cm (Two tailed test)

(iii) Level of significance: $\alpha = 0.05$

(iv) Test statistic: $Z = \dfrac{\bar{x} - \mu}{\left(\dfrac{\sigma}{\sqrt{n}}\right)} = \dfrac{3.4 - 3.25}{\left(\dfrac{2.61}{\sqrt{900}}\right)} = 1.72$

$$|Z| = 1.72$$

(v) Critical value: $|Z_{0.05}| = 1.96$

(vi) Decision: Since $|Z| < |Z_{0.05}|$, the null hypothesis is accepted at 5% level of significance i.e., the sample has been drawn from the population with mean $\mu = 3.25$ cm.

95% fiducial limits:

$$\bar{x} \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right) = 3.4 \pm 1.96\left(\frac{2.61}{\sqrt{900}}\right) = 3.4 \pm 0.1705,$$

i.e., 3.5705 and 3.2295

Example 5

A type company claims that the lives of tyres have mean 42000 *km with s.d. of* 4000 *km. A change in the production process is believed to result in better product. A test sample of* 81 *new tyres has a mean life of* 42500 *km. Test at 5% level of significance that the new product is significantly better than the old one.*

Solution

$$n = 81, \quad \bar{x} = 42500 \text{ km}, \quad \mu = 42000 \text{ km}, \quad \sigma = 4000 \text{ km}$$

(i) Null Hypothesis H_0: $\mu = 42000$ km, i.e., the new product is not significantly better than the old one.

(ii) Alternative Hypothesis H_1: $\mu > 42000$ km (Right tailed test)

(iii) Level of significance: $\alpha = 0.05$

(iv) Test statistic: $Z = \dfrac{\bar{x} - \mu}{\left(\dfrac{\sigma}{\sqrt{n}}\right)} = \dfrac{42500 - 42000}{\left(\dfrac{4000}{\sqrt{81}}\right)} = 1.125$

$$|Z| = 1.125$$

(v) Critical value: $Z_{0.05} = 1.645$

(vi) Decision: Since $|Z| < Z_{0.05}$, the null hypothesis is accepted at 5% level of significance, i.e., the new product is not significantly better than the old one.

Example 6

The mean breaking strength of cables supplied by a manufacturer is 1800 *with standard deviation* 100. *By a new technique in the manufacturing process it is claimed that the breaking strength of the cable has increased. In order to test the claim a sample of* 50 *cables is tested. It is found that the mean breaking strength is* 1850. *Can we support the claim at 1% level of significance?*

Solution

$$n = 50, \quad \bar{x} = 1850, \quad \mu = 1800, \quad \sigma = 100$$

(i) Null Hypothesis H_0: $\mu = 1800$, i.e., the mean breaking strength of cables supplied by manufacturer is 1800.

(ii) Alternative Hypothesis H_1: $\mu > 1800$ (Right tailed test)

(iii) Level of significance: $\alpha = 0.01$

(iv) Test statistic: $Z = \dfrac{\bar{x} - \mu}{\left(\dfrac{\sigma}{\sqrt{n}}\right)} = \dfrac{1850 - 1800}{\left(\dfrac{100}{\sqrt{50}}\right)} = 3.54$

$$|Z| = 3.54$$

(v) Critical value: $Z_{0.01} = 2.33$

(vi) Decision: Since $|Z| > Z_{0.01}$, the null hypothesis is rejected at 1% level of significance, i.e., the mean breaking strength of cables supplied is more than 1800.

Example 7

An ambulance service claims that it takes on the average 10 *minutes to reach its destination in emergency calls. A sample of* 36 *calls has a* *mean of 11 minutes and the variance of 16 minutes. Test the claim at 0.05 level of significance.*

Solution

$$n = 36, \quad \bar{x} = 11 \text{ minutes}, \quad \mu = 10 \text{ minutes}, \quad s = \sqrt{16} = 4 \text{ minutes}$$

(i) Null Hypothesis H_0: $\mu = 10$ minutes, i.e., ambulance service takes 10 minutes to reach the destination.

(ii) Alternative Hypothesis H_1: $\mu > 10$ minutes (Right tailed test)

(iii) Level of significance: $\alpha = 0.05$

(iv) Test statistic: $$Z = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)} = \frac{11 - 10}{\left(\frac{4}{\sqrt{36}}\right)} = 1.5$$

$$|Z| = 1.5$$

(v) Critical value: $Z_{0.05} = 1.645$

(vi) Decision: Since $|Z| < Z_{0.05}$, the null hypothesis is accepted at 5% level of confidence, i.e., the ambulance service takes on the average 10 minutes to reach its destination.

6.8 TEST OF SIGNIFICANCE FOR DIFFERENCE OF MEANS – LARGE SAMPLES

Let $\bar{x}_1$ and $\bar{x}_2$ be the sample means of two independent large random samples with sizes n_1 and n_2 ($n_1 > 30$, $n_2 > 30$) drawn from two populations with means μ_1 and μ_2 and standard deviations σ_1 and σ_2.

Working Rule

(i) Null Hypothesis H_0: $\mu_1 = \mu_2$, i.e., the two samples have been drawn from two different populations having the same means and equal standard deviations.

(ii) Alternative Hypothesis H_1: $\mu_1 \neq \mu_2$ (two tailed test)
or H_1: $\mu_1 < \mu_2$ (one tailed test)
or H_1: $\mu_1 > \mu_2$ (one tailed test)

(iii) Level of significance: Select the level of significance α.

(iv) Test statistic: There are two cases for calculating test statistic.

(a) When the population standard deviations σ_1 and σ_2 are known

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

(b) When the population standard deviations σ_1 and σ_2 are not known

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where s_1 and s_2 are sample standard deviations.

(v) Critical Value: Find the critical value (tabulated value) Z_α of Z at the given level of significance.

(vi) Decision: If $|Z| < Z_\alpha$ at the level of significance α, the null hypothesis is accepted. If $|Z| > Z_\alpha$ at the level of significance α, the null hypothesis is rejected.

Note:

1. Null Hypothesis H_0 is rejected when $|Z| > 3$ without mentioning any level of significance.
2. Confidence limits:
   (i) 95% confidence limits $= (\bar{x}_1 - \bar{x}_2) \pm 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
   (ii) 99% confidence limits $= (\bar{x}_1 - \bar{x}_2) \pm 2.58\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

If population standard deviation σ_1 and σ_2 are not known, s_1 and s_2 are used in equations.

Example 1

Test the significance of the difference between the means of two normal population with the same standard deviation from the following data:

	Size	*Mean*	*SD*
Sample I	100	64	6
Sample II	200	67	8

Solution

$$n_1 = 100, \quad n_2 = 200, \quad \bar{x}_1 = 64, \quad \bar{x}_2 = 67 \quad s_1 = 6, \quad s_2 = 8$$

(i) Null Hypothesis $H_0 : \mu_1 = \mu_2$ i.e., there is no significant difference between two means.

(ii) Alternative Hypothesis H_1: $\mu_1 \neq \mu_2$ (Two tailed test)

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: $$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{64 - 67}{\sqrt{\frac{(6)^2}{200} + \frac{(8)^2}{100}}} = -3.31$$

$$|Z| = 3.31$$

(v) Critical value: $|Z_{0.05}| = 1.96$

(vi) Decision: Since $|Z| > |Z_{0.05}|$, the null hypothesis is rejected at 5% level of significance, i.e., the samples do not support the hypothesis that the two population have the same mean although they may have the same standard deviation.

Example 2

The means of simple samples of sizes 1000 and 2000 are 67.5 and 68 cm respectively. Can the samples be regarded as drawn from the same population of S.D. 2.5 cm.

Solution

$$n_1 = 1000,\ n_2 = 2000,\ \bar{x}_1 = 67.5 \text{ cm},\ \bar{x}_2 = 68 \text{ cm},\ \sigma = 2.5 \text{ cm}$$

(i) Null Hypothesis H_0: $\mu_1 = \mu_2$ i.e., the samples have been drawn from the same population of S.D. 2.5 cm

(ii) Alternative Hypothesis $H_1 : \mu_1 \neq \mu_2$ (Two tailed test)

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: $Z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma^2}{n_1} + \dfrac{\sigma^2}{n_2}}} = \dfrac{67.5 - 68}{\sqrt{\dfrac{(2.5)^2}{1000} + \dfrac{(2.5)^2}{2000}}} = -5.16$

$$|Z| = 5.16$$

(v) Critical value: $|Z_{0.05}| = 1.96$

(iv) Decision: Since $|Z| > |Z_{0.05}|$, the null hypothesis is rejected at 5% level of significance, i.e., the samples cannot be regarded as drawn from the same population of SD 2.5 cm.

Example 3

The mean life of a sample of 10 electric bulbs was found to be 1456 hours with SD of 423 hours. A second sample of 17 bulbs chosen from a different batch showed a mean life of 1280 with SD of 398 hours. Is there a significant difference between the means of two batches?

Solution

$$n_1 = 10, n_2 = 17, \bar{x}_1 = 1456 \text{ hours}, \bar{x}_2 = 1280 \text{ hours}, s_1 = 423 \text{ hours}, s_2 = 398 \text{ hours}$$

(i) Null Hypothesis H_0: $\mu_1 = \mu_2$, i.e., there is no significant difference between the means of two batches.

(ii) Alternative Hypothesis H_1: $\mu_1 \neq \mu_2$ (Two tailed test)

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: $Z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \dfrac{1456 - 1280}{\sqrt{\dfrac{(423)^2}{10} + \dfrac{(398)^2}{17}}} = 1.07$

$$|Z| = 1.07$$

(v) Critical value: $|Z_{0.05}| = 1.96$

(vi) Decision: Since $Z < |Z_{0.05}|$, the null hypothesis is accepted at 5% level of significance, i.e., there is no significant difference between the means of two batches.

Example 4

The average of marks scored by 32 boys is 72 with standard deviation 8 while that of 36 girls is 70 with standard deviation 6. Test at 1% level of significance whether the boys perform better than the girls.

Solution

$$n_1 = 32,\ n_2 = 36,\ \bar{x}_1 = 72,\ \bar{x}_2 = 70,\ s_1 = 8,\ s_2 = 6$$

(i) Null Hypothesis H_0: $\mu_1 = \mu_2$, i.e., there is no significant difference between the performance of boys and girls.

(ii) Alternative Hypothesis H_1: $\mu_1 > \mu_2$ (Right tailed test)

(iii) Level of significance: $\alpha = 0.01$

(iv) Test statistic: $Z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \dfrac{72 - 70}{\sqrt{\dfrac{(8)^2}{32} + \dfrac{(6)^2}{36}}} = 1.1547$

$$|Z| = 1.1547$$

(v) Critical value: $Z_{0.01} = 2.33$

(vi) Decision: Since $|Z| < Z_{0.01}$, the null hypothesis is accepted at 1% level of significance, i.e., the boys do not perform better than the girls.

Example 5

A simple sample of heights of 6400 English men has a mean of 170 cm and a s.d. of 6.4 cm, while a simple sample of heights of 1600 Americans has a mean of 172 cm and a s.d. of 63 cm. Do the data indicate that American are, on the average, taller than the English men?

Solution

$$n_1 = 1600,\ n_2 = 6400,\ \bar{x}_1 = 172 \text{ cm},\ \bar{x}_2 = 170 \text{ cm},\ s_1 = 6.3 \text{ cm},\ s_2 = 6.4 \text{ cm}$$

(i) Null Hypothesis H_0: $\mu_1 = \mu_2$, i.e., there is no significant difference in heights of Americans and English men.

(ii) Alternative Hypothesis H_1: $\mu_1 > \mu_2$ (Right tailed test)

(iii) Level of significance: $\alpha = 0.01$ (assumption)

(iv) Test statistic: $$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{172 - 170}{\sqrt{\frac{(6.3)^2}{1600} + \frac{(6.4)^2}{6400}}} = 11.32$$

$$|Z| = 11.32$$

(v) Critical value: $Z_{0.01} = 2.33$

(vi) Decision: Since $|Z| > Z_{0.01}$, the null hypothesis is rejected at 1% level of significance, i.e., Americans are, on the average, taller than English men.

Example 6

In a certain factory there are two different processes of manufacturing the same item. The average weight in a sample of 250 items produced from one process is found to be 120 gm with a s.d. of 12 gm; the corresponding figures in a sample of 400 items from the other process are 124 gm and 14 gm. Is this difference between the two sample means significant?

Solution

$n_1 = 250,\ n_2 = 400,\ \bar{x}_1 = 120 \text{ gm},\ \bar{x}_2 = 124 \text{ gm},\ s_1 = 12 \text{ gm},\ s_2 = 14 \text{ gm}$

(i) Null Hypothesis H_0: $\mu_1 = \mu_2$, i.e., there is no significant difference between the two sample means.

(ii) Alternative Hypothesis H_1: $\mu_1 \neq \mu_2$ (Two tailed test)

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: $$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{120 - 124}{\sqrt{\frac{(12)^2}{250} + \frac{(14)^2}{400}}} = -3.87$$

$$|Z| = 3.87$$

(v) Critical value: $|Z_{0.05}| = 1.96$

(vi) Decision: Since $|Z| > |Z_{0.05}|$, the null hypothesis is rejected at 5% level of significance, i.e., there is significant difference between two sample means.

Example 7

The mean height of 50 male students who participate in sports is 68.2 inches with a s.d. of 2.5 inches. The mean height of 50 male students who have not participated in sport is 67.2 inches with a s.d. of 2.8 inches. Test the hypothesis that the height of students who have participated in sports is more than the students who have not participated in sports.

Solution

$n_1 = 50,\ n_2 = 50,\ \bar{x}_1 = 68.2 \text{ inch},\ \bar{x}_2 = 67.2 \text{ inch},\ s_1 = 2.5 \text{ inch},\ s_2 = 2.8 \text{ inch}$

(i) Null Hypothesis H_0: $\mu_1 = \mu_2$, i.e., there is no significant difference in heights of students who have participated in sports or not.

(ii) Alternative Hypothesis H_1: $\mu_1 > \mu_2$ (Right tailed test)

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: $$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{68.2 - 67.2}{\sqrt{\frac{(2.5)^2}{50} + \frac{(2.8)^2}{50}}} = 1.88$$

$$|Z| = 1.88$$

(v) Critical value: $Z_{0.05} = 1.645$

(vi) Decision: Since $|Z| > Z_{0.05}$, the null hypothesis is rejected at 5% level of significance, i.e., the height of students who have participated in sports is more than the students who have not participated in sports.

6.9 TEST OF SIGNIFICANCE FOR DIFFERENCE OF STANDARD DEVIATIONS – LARGE SAMPLES

Let s_1 and s_2 be the standard deviations of two independent large random samples with sizes n_1 and n_2 ($n_1 > 30, n_2 > 30$) drawn from two populations with standard deviations σ_1 and σ_2.

Working Rule

(i) Null Hypothesis H_0: $\sigma_1 = \sigma_2$, i.e., the two samples have been drawn from two different populations having same standard deviations.

(ii) Alternative Hypothesis H_1: $\sigma_1 \neq \sigma_2$ (Two tailed test)

or H_1: $\sigma_1 < \sigma_2$ (One tailed test)

or H_1: $\sigma_1 > \sigma_2$ (One tailed test)

(iii) Level of significance: Select the level of significance α.

(iv) Test statistic: There are two cases for calculating test statistic.

(a) When the population standard deviations σ_1 and σ_2 are known

$$Z = \frac{s_1 - s_2}{\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}}$$

(a) When the population standard deviations σ_1 and σ_2 are not known

$$Z = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}}$$

where s_1 and s_2 are sample standard deviation.

(v) Critical value: Find the critical vale (tabulated value) Z_α of Z at the given level of significance.

(vi) Decision: If $|Z| < Z_\alpha$ at the level of significance α, the null hypothesis is accepted. If $|Z| > Z_\alpha$ at the level of significance α, the null hypothesis is rejected.

Example 1

The SD of a random sample of 1000 is found to be 2.6 and the SD of another random sample of 500 is 2.7. Assuming the samples to be independent, find whether the two samples could have come from populations with the same SD.

Solution

$$n_1 = 1000,\ n_2 = 500,\ s_1 = 2.6,\ s_2 = 2.7$$

(i) Null Hypothesis H_0: $\sigma_1 = \sigma_2$, i.e., there is no significant difference between two standard deviations.

(ii) Alternative Hypothesis H_1: $\sigma_1 \neq \sigma_2$ (Two tailed test)

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: $Z = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}} = \frac{2.6 - 2.7}{\sqrt{\frac{(2.6)^2}{2(1000)} + \frac{(2.7)^2}{2(500)}}} = -0.97$

$$|Z| = 0.97$$

(v) Critical value: $|Z_{0.05}| = 1.96$

(vi) Decision: Since $|Z| < |Z_{0.05}|$, the null hypothesis H_0 is accepted at 5% level of significance, i.e., there is no significance difference between two standard deviations and the two samples could have come from populations with the same SD.

Example 2

Random samples drawn from two countries gave the following data relating to the heights of adult males:

	Country A	*Country B*
Standard deviation (in inches)	2.58	2.50
Number in samples	1000	1200

Is the difference between the standard deviations significant?

Solution

$$n_1 = 1000,\ n_2 = 1200,\ s_1 = 2.58 \text{ inch},\ s_2 = 2.50 \text{ inch}$$

(i) Null Hypothesis H_0: $\sigma_1 = \sigma_2$, i.e., there is no significant difference between two standard deviations.

(ii) Alternative Hypothesis H_1: $\sigma_1 \neq \sigma_2$ (Two tailed test)

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: $Z = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}} = \frac{2.58 - 2.50}{\sqrt{\frac{(2.58)^2}{2(1000)} + \frac{(2.50)^2}{2(1200)}}} = 0.077$

$$|Z| = 0.077$$

(v) Critical value: $|Z_{0.05}| = 1.96$

(vi) Decision: Since $|Z| < |Z_{0.05}|$, the null hypothesis is accepted at 5% level of significance, i.e., there is no significance difference between the standard deviations.

Example 3

Examine whether the two samples for which the data are given in the following table could have been drawn from populations with the same SD.

	Size	*SD*
Sample I	100	5
Sample II	200	7

Solution

$$n_1 = 100,\ n_2 = 200,\ s_1 = 5,\ s_2 = 7$$

(i) Null Hypothesis H_0: $\sigma_1 = \sigma_2$, i.e., the two samples could have been drawn from populations with the same SD.

(ii) Alternative Hypothesis H_1: $\sigma_1 \neq \sigma_2$ (Two tailed test)

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: $$Z = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}} = \frac{5-7}{\sqrt{\frac{(5)^2}{2(100)} + \frac{(7)^2}{2(200)}}} = -4.02$$

$$|Z| = 4.02$$

(v) Critical value: $|Z_{0.05}| = 1.96$

(vi) Decision: Since $|Z| > |Z_{0.05}|$, the null hypothesis is rejected at 5% level of significance, i.e., the two samples could not have been drawn from populations with the same SD.

EXERCISE 6.2

1. A random sample of 100 students gave a mean weight of 58 kg with a SD of 4 kg. Test the hypothesis that the mean weight in the population is 60 kg.
   [Ans.: The mean weight in the population is not 60 kg]
2. A sample of 400 items is taken from a normal population whose mean is 4 and whose variance is also 4. If the sample mean is 4.45, can the sample be regarded as truly random sample?
   [Ans.: Sample cannot be regarded as truly random sample]
3. The mean IQ of a sample of 1600 children was 99. Is it likely that this was a random sample from a population with mean IQ 100 and SD 15?
   [Ans.: Sample was not drawn from a population with mean 100 and SD 15]
4. In a random sample of 60 workers, the average time taken by them to get to work is 33.8 minutes with a standard deviation of 6.1 minutes. Can we reject the null hypothesis $\mu = 32.6$ minutes in favour of alternative hypothesis $\mu > 32.6$ at $\alpha = 0.025$ level of significance
   [Ans.: The null hypothesis is accepted]
5. It is claimed that a random sample of 49 types has a mean life of 15200 km. This sample was drawn from a population whose mean is 15150 km and a standard deviation of 1200 km. Test the significance at 0.05 level.
   [Ans.: The null hypothesis is accepted]

6. An ambulance service claims that it takes on the average less than 10 minutes to reach its destination in emergency calls. A sample of 36 calls has a mean of 11 minutes and the variance of 16 minutes. Test the claim at 0.05 level of significance.
   [Ans.: The null hypothesis is accepted]
7. Samples of students were drawn from two universities and from their weights in kilograms, the mean and standard deviations are calculated. Make a large sample test to test the significance of the difference between the means.

	Mean	*SD*	*Size of the sample*
University A	55	10	400
University B	57	15	100

[Ans.: There is no significant difference between the means]

8. A researchers wants to know the intelligence of students in a school. He selected two groups of students. In the first group, there are 150 students having mean IQ of 75 with a SD of 15. In the second group there are 250 students having mean IQ of 70 with SD of 20. Test the significance that the groups have come from same population.
   [Ans.: The groups have not come from same population]
9. Random samples drawn from two places gave the following data relating to the heights of children:

	Mean height in cm	*SD in cm*	*No. of children in sample*
Place A	68.50	2.5	1200
Place B	68.58	3.0	1500

Test at 5% level of significance that the mean height is the same for children at two places.

[Ans.: The mean height is same for children at two places]

10. The mean life of a sample of 10 electric bulbs was found to be 1456 hours with SD of 423 hours. A second sample of 17 bulbs chosen from a different batch showed a mean life of 1280 hours with SD of 398 hours. Is there a significant difference between the means of two batches?
    [Ans.: There is no difference between the mean life of two batches]
11. The SD of a random sample of 900 members is 4.6 and that of another independent sample of 1600 members is 4.8. Examine if the two samples could have been drawn from a population with SD 4?
    [Ans.: Two samples could have been drawn from a population with SD 4]

12. The variability of two sets of plots is as given below:

	Set of 40 plots	Set of 60 Plots
SD per plot	34 kg	28 kg

Examine whether the difference in the variability in yields is significant.

[Ans.: The difference in the variability in yields is significant]

6.10 SMALL SAMPLE TESTS

If the samples are large ($n > 30$) then the sampling distribution of a statistic is normal. But if the samples are small ($n \leq 30$) then the above result does not hold good. For estimation of the parameter as well as for testing a hypothesis, following distributions are used:

(i) Student's t-distribution
(ii) Snedecer's F-distribution
(iii) Chi-square (χ^2) distribution

6.11 STUDENT'S *t*-DISTRIBUTION

The theory of small or exact sample was developed by Irish statistician William S. Gosset who used to write under pen-name of student. The quantity t is defined as

$$t = \frac{\text{Difference of population parameter and the corresponding statistic}}{\text{Standard error of statistic}}$$

with $(n-1)$ degrees of freedom if the sample size is n.

Let $x_1, x_2, \ldots, x_n$ be a random sample of size n ($n \leq 30$) drawn from a normal population with mean μ and SD σ. The student's t statistic is defined by

$$t = \frac{\bar{x}-\mu}{\left(\frac{\mu}{\sqrt{n}}\right)} \quad \text{or} \quad t = \frac{\bar{x}-\mu}{\left(\frac{s}{\sqrt{n-1}}\right)}$$

where $\bar{x}$ is sample mean and $s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$ is an unbiased estimate of σ^2. The test statistic $t = \frac{\bar{x}-\mu}{\left(\frac{s}{\sqrt{n-1}}\right)}$ is a random variable having t-distribution with $v = n-1$ degrees of freedom and with probability density function $f(t) = c\left(1+\frac{t^2}{v}\right)^{\frac{-(v+1)}{2}}$, where $v = n-1$ and c is a constant required to make the area under the curve unity, i.e., $\int_{-\infty}^{\infty} f(t)dt = 1$.

The t-distribution is used when (i) the sample size is less than or equal to 30, and (ii) population standard deviation is not known.

6.11.1 Assumptions for *t*-test

(i) Samples are drawn from normal population and are random.
(ii) The population standard deviation may not be known.
(iii) For testing the equality of two population mean, the population variances are regarded as equal.
(iv) In case of two samples, some adjustments in degrees of freedom for t are made.

6.11.2 Properties of *t*-distribution

(i) The t-distribution is asymptotic to the x-axis, i.e., it extends to infinity on either side.
(ii) The t-distribution is symmetrical about the mean.
(iii) The shape of the curve varies with the degrees of freedom.
(iv) The larger the number of degrees of freedom, the more closely t-distribution resembles standard normal distribution.
(v) Sampling distribution of t does not depend on population parameter but it depends only on degree of freedom v, i.e., on the sample size.

6.11.3 Applications of *t*-distribution

The t-distribution has following applications in testing of hypotheses for small samples:

(i) To test the significance of the sample mean, when the population variance σ is not known
(ii) To test the significance of the mean of the sample i.e., to test if the sample mean differs significantly from the population mean
(iii) To test the significance of the difference between two sample means, the population variances being equal and unknown
(iv) To test the significance of an observed sample correlation coefficient

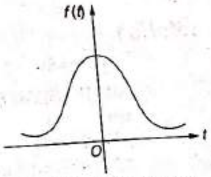


Fig. 6.5 t-distribution curve

6.12 *t*-TEST: TEST OF SIGNIFICANCE FOR SINGLE MEAN

If $x_1, x_2, \ldots, x_n$ is a random sample of size n ($n \leq 30$) drawn from a normal population with mean μ and SD σ and if the sample mean $\bar{x}$ differs significantly from the population mean μ then the student's t statistic is given by

$$t = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n-1}}\right)}, \quad \text{where} \quad s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}} \quad \text{with } \nu = n - 1$$

Note: Confidence Limit

(i) 95% confidence limits = $\bar{x} \pm t_{0.05}\left(\frac{s}{\sqrt{n-1}}\right)$

where $t_{0.05}$ is the 5% critical value of t for $\nu = n - 1$ degree of freedom for a Two tailed test.

(ii) 99% confidence limits = $\bar{x} \pm t_{0.01}\left(\frac{s}{\sqrt{n-1}}\right)$

where $t_{0.01}$ is the 1% critical value of t for $\nu = n - 1$ degree of freedom for a Two tailed test.

Example 1

A machinist is making engine parts with axle diameter of 0.7 *cm. A random sample of* 10 *parts shows a mean diameter of* 0.742 *cm with a standard deviation of* 0.04 *cm. Compute the statistic you would use to test whether work is meeting the specification at* 0.05 *level of significance.*

Solution

$$n = 10, \quad \bar{x} = 0.742 \text{ cm}, \quad s = 0.04 \text{ cm} \quad \mu = 0.7 \text{ cm}$$

(i) Null Hypothesis H_0: $\mu = 0.7$ cm, i.e., the product is meeting the specification.

(ii) Alternative Hypothesis H_1: $\mu \neq 0.7$ cm (Two tailed test)

(iii) Level of significance: $\alpha = 0.05$

(iv) Test statistic: $t = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n-1}}\right)} = \frac{0.742 - 0.7}{\left(\frac{0.04}{\sqrt{10-1}}\right)} = 3.15$

$$|t| = 3.15$$

(v) Critical value: $\nu = n - 1 = 10 - 1 = 9$

$$t_{0.05}(\nu = 9) = 2.262$$

(vi) Decision: Since $|t| > t_{0.05}$, the null hypothesis is rejected at 5% level of significance i.e., the product is not meeting the specification.

Example 2

Ten objects are chosen at random from a large population and their weights are found to be in grams: 63, 63, 64, 65, 66, 69, 69, 70, 70, 71. *Discuss the suggestion that the mean weight is* 65 *g.*

Solution

$$n = 10, \mu = 65 \text{ g}$$

$$\left.\begin{aligned} \bar{x} &= 67 \text{ g} \\ s &= 2.966 \text{ g} \end{aligned}\right\} \text{From calculator}$$

(i) Null Hypothesis H_0: $\mu = 65$ g, i.e., there is no significant difference in the mean weight of sample and population.

(ii) Alternate Hypothesis H_1: $\mu \neq 65$ g (Two tailed test)

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: $t = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n-1}}\right)} = \frac{67 - 65}{\left(\frac{2.966}{\sqrt{10-1}}\right)} = 2.023$

$$|t| = 2.023$$

(v) Critical value: $\nu = n - 1 = 10 - 1 = 9$

$$t_{0.05}(\nu = 9) = 2.262$$

(vi) Decision: Since $|t| < t_{0.05}$, the null hypothesis is accepted at 5% level of significance, i.e., the mean weight is 65 g.

Example 3

The mean lifetime of a sample of 25 *bulbs is found as* 1550 *hours with a SD of* 120 *hours. The company manufacturing the bulbs claims that the average life of their bulbs is* 1600 *hours. Is the claim acceptance at* 5% *level of significance?*

Solution

$$n = 25, \quad \bar{x} = 1550 \text{ hours}, \quad s = 120 \text{ hours}, \quad \mu = 1600 \text{ hours}$$

(i) Null Hypothesis H_0: $\mu = 1600$ hours, i.e., the average life of bulbs is 1600 hours.

(ii) Alternative Hypothesis H_1: $\mu < 1600$ hours (One tailed test)

(iii) Level of significance: $\alpha = 0.05$

(v) Critical value: $\nu = n - 1 = 10 - 1 = 9$

$$t_{0.05}\ (\nu = 9) = 2.262$$

(vi) Decision: Since $|t| < t_{0.05}$, the null hypothesis is accepted at 5% level of significance, i.e., population has mean IQ of 100.

$$\begin{aligned}95\%\text{ confidence limits} &= \bar{x} \pm t_{0.05}\left(\frac{s}{\sqrt{n-1}}\right)\\ &= 97.2 \pm 2.262\left(\frac{13.54}{\sqrt{10-1}}\right)\\ &= 97.2 \pm 10.21\\ &= 87 \text{ and } 107.41\end{aligned}$$

Example 7

The heights of 10 *males of a given locality are found to be* 175, 168, 155, 170, 152, 170, 175, 160, 160 *and* 165 *cm. Based on this sample, find the 95% confidence limits for the heights of males in that locality.*

Solution

$$n = 10$$

$$\left.\begin{aligned}\bar{x} &= 165\\ s &= 7.6\end{aligned}\right\}\ \text{From calculator}$$

$$\nu = n - 1 = 10 - 1 = 9$$

From t-table

$$t_{0.05}(\nu = 9) = 2.262 \text{ (Two tailed test)}$$

The 95% confidence limits for μ are

$$\left[\bar{x} - t_{0.05}\left(\frac{s}{\sqrt{n-1}}\right),\ \bar{x} + t_{0.05}\left(\frac{s}{\sqrt{n-1}}\right)\right]$$

i.e.,
$$\left[165 - \frac{2.262(7.6)}{\sqrt{10-1}},\ 165 + \frac{2.262(7.6)}{\sqrt{10-1}}\right]$$

i.e., $[159.27,\ 170.73]$

i.e., the heights of males in the locality are likely to be in limits 159.27 cm and 170.73 cm.

6.13 *t*-TEST: TEST OF SIGNIFICANCE FOR DIFFERENCE OF MEANS

Let $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ be two independent samples of sizes n_1 and n_2 ($n_1 \le 30$, $n_2 \le 30$) with means $\bar{x}$ and $\bar{y}$ and standard deviations s_1 and s_2 from a normal population with means μ_1 and μ_2 and same standard deviations. The student's t statistic is given by

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \quad \text{with } \nu = n_1 + n_2 - 2$$

where
$$\bar{x} = \frac{\Sigma \bar{x}}{n_1}$$

$$\bar{y} = \frac{\Sigma \bar{y}}{n_2}$$

and
$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2 + \Sigma(y-\bar{y})^2}{n_1 + n_2 - 2}}$$

In terms of standard deviations s_1 and s_2.

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

where
$$s = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}$$

and
$$s_1 = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n_1}}$$

$$s_2 = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n_2}}$$

Note

1. If $n_1 = n_2 = n$ and the samples are independent, i.e., the observations in the two samples are not all related then test statistic is given by

$$t = \frac{x - y}{\sqrt{\dfrac{s_1^2 + s_2^2}{n-1}}} \quad \text{with } \nu = 2n - 2$$

2. If $n_1 = n_2 = n$ and if the pairs of values of x and y are associated or correlated in some way (or not independent), the above formula for testing of hypothesis cannot be used.
   Let $d_i = x_i - y_i$ denote the difference (with proper sign) in the values of x and y for the ith pair ($i = 1, 2, \ldots, n$).

   The test statistics is given by

$$t = \frac{\bar{d}}{\left(\dfrac{s}{\sqrt{n-1}}\right)} \quad \text{with } \nu = n - 1$$

where $\bar{d}$ and s denote the mean and standard deviation of the difference d_i respectively, i.e.,

$$\bar{d} = \frac{\sum d_i}{n}$$

$$s = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n}} = \frac{\sum d_i^2}{n} - \left(\frac{\sum d_i}{n}\right)^2$$

(3) Confidence Limits

(i) 95% confidence limits $= (\bar{x} - \bar{y}) \pm t_{0.05}\left(\dfrac{1}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}\right)$

where $t_{0.05}$ is the 5% critical value of t for $v = n_1 + n_2 - 2$ degree of freedom for a Two tailed test.

(ii) 99% confidence limits $= (\bar{x} - \bar{y}) \pm t_{0.01}\left(\dfrac{1}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}\right)$

where $t_{0.01}$ is the 1% critical value of t for $v = n_1 + n_2 - 2$ degree of freedom for a Two tailed test.

Example 1

The means of two random samples of size 9 *and* 7 *are* 196.42 *and* 198.82 *respectively. The sum of squares of the deviation from the mean are* 26.94 *and* 18.73 *respectively. Can the sample be considered to have been drawn from the same population?*

Solution

$$n_1 = 9,\ n_2 = 7,\ \bar{x} = 196.42,\ \bar{y} = 198.82$$

$$\sum(x - \bar{x})^2 = 26.94,\ \sum(y - \bar{y})^2 = 18.73$$

$$s = \sqrt{\frac{\sum(x - \bar{x})^2 + \sum(y - \bar{y})^2}{n_1 + n_2 - 2}} = \sqrt{\frac{26.94 + 18.73}{9 + 7 - 2}} = 1.806$$

(i) Null Hypothesis H_0: $\mu_1 = \mu_2$, i.e., the samples are drawn from the same population.

(ii) Alternative Hypothesis H_1: $\mu_1 \neq \mu_2$ (Two tailed test)

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: $t = \dfrac{\bar{x} - \bar{y}}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} = \dfrac{196.42 - 198.82}{1.8061\sqrt{\dfrac{1}{9} + \dfrac{1}{7}}} = -2.6368$

$$|t| = 2.6368$$

(v) Critical value: $v = n_1 + n_2 - 2 = 9 + 7 - 2 = 14$

$$t_{0.05}\,(v = 14) = 2.145$$

(vi) Decision: Since $|t| > t_{0.05}$, the null hypothesis is rejected at 5% level of significance, i.e., the samples are not drawn from the same population.

Example 2

Samples of two types of electric bulbs were tested for length of life and the following data were obtained.

	Size	*Mean*	*SD*
Sample 1	8	1234 hr	36 hr
Sample 2	7	1036 hr	40 hr

Is the difference in the means sufficient to warrant that type 1 *bulbs are superior to type* 2 *bulbs?*

Solution

$$n_1 = 8,\ n_2 = 7,\ \bar{x}_1 = 1234 \text{ hr},\ \bar{x}_2 = 1036 \text{ hr}$$

$$s_1 = 36 \text{ hr},\ s_2 = 40 \text{ hr}$$

$$s = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(8)(36)^2 + (7)(40)^2}{8 + 7 - 2}} = 40.73 \text{ hr}$$

(i) Null Hypothesis H_0: $\mu_1 = \mu_2$, i.e., the type 1 bulbs are not superior to type 2 bulbs.

(ii) Alternative Hypothesis H_1: $\mu_1 > \mu_2$ (One tailed test)

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: $t = \dfrac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} = \dfrac{1234 - 1036}{40.73\sqrt{\dfrac{1}{8} + \dfrac{1}{7}}} = 9.39$

$$|t| = 9.39$$

(v) Critical value: $v = n_1 + n_2 - 2 = 8 + 7 - 2 = 13$

$$t_{0.05}\,(v = 13) = 1.771$$

(vi) Decision: Since $|t| > t_{0.05}$, the null hypothesis is rejected at 5% level of significance, i.e., the type 1 bulbs are superior to type 2 bulbs.

Example 3

The mean height and SD height of 8 randomly chosen soldiers are 166.9 cm and 8.29 cm respectively. The corresponding values of 6 randomly chosen sailors are 170.3 cm and 8.50 cm respectively. Based on this data, can we conclude that soldiers are, in general, shorter than sailors?

Solution

$$n_1 = 8,\ n_2 = 6,\ \bar{x}_1 = 166.9 \text{ cm},\ \bar{x}_2 = 170.3 \text{ cm}$$

$$s_1 = 8.29 \text{ cm},\ s_2 = 8.50 \text{ cm}$$

$$s = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(8)(8.29)^2 + (6)(8.50)^2}{8+6-2}} = 9.05 \text{ cm}$$

(i) Null Hypothesis H_0: $\mu_1 = \mu_2$, i.e., there is no significant difference between the heights of soldiers and sailors.

(ii) Alternative Hypothesis H_1: $\mu_1 < \mu_2$ (One tailed test)

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: $t = \dfrac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} = \dfrac{166.9 - 170.3}{9.05\sqrt{\dfrac{1}{8} + \dfrac{1}{6}}} = -0.696$

$$|t| = 0.696$$

(v) Critical value: $\nu = n_1 + n_2 - 2 = 8 + 6 - 2 = 12$

$$t_{0.05}\ (\nu = 12) = 1.782$$

(vi) Decision: Since $|t| < t_{0.05}$, the null hypothesis is accepted at 5% level of significance, i.e., there is no significant difference between the heights of soldiers and sailors and we cannot conclude that sailors are, in general, shorter than sailors.

Example 4

Two types of batteries are tested for their length of life and the following data are obtained:

	No. of Samples	*Mean life in hours*	*Variance*
Type A	9	600	121
Type B	8	640	144

Is there a significant difference in the two means? Find 95% confidence limits for the difference in means.

Solution

$$n_1 = 9,\ n_2 = 8,\ \bar{x}_1 = 600 \text{ hours},\ \bar{x}_2 = 640 \text{ hours}$$

$$s_1 = \sqrt{121} = 11 \text{ hours},\ s_2 = \sqrt{144} = 12 \text{ hours}$$

$$s = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(9)(121) + (8)(144)}{9+8-2}} = 12.22 \text{ hours}$$

(i) Null Hypothesis H_0: $\mu_1 = \mu_2$, i.e., there is no significant difference in two means.

(ii) Alternative Hypothesis H_1: $\mu_1 \neq \mu_2$ (Two tailed test)

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: $t = \dfrac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} = \dfrac{600 - 640}{12.22\sqrt{\dfrac{1}{9} + \dfrac{1}{8}}} = -6.74$

$$|t| = 6.74$$

(v) Critical value: $\nu = n_1 + n_2 - 2 = 9 + 8 - 2 = 15$

$$t_{0.05}\ (\nu = 15) = 2.132$$

(vi) Decision: Since $|t| > t_{0.05}$, the null hypothesis is rejected at 5% level of confidence, i.e., there is significant difference in the two means.

$$\begin{aligned} 95\% \text{ confidence limits for } (\mu_1 - \mu_2) &= (\bar{x}_1 - \bar{x}_2) \pm t_{0.05}\left(s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) \\ &= (600 - 640) \pm 2.132\left(12.22\sqrt{\frac{1}{9} + \frac{1}{8}}\right) \\ &= -40 \pm 12.66 \\ &= -27.34 \text{ and } -52.66 \end{aligned}$$

Example 5

A group of 5 patients treated with medicine A weigh 42, 39, 48, 60 and 41 kg. Second group of 7 patients from the same hospital treated with *medicine B weigh* 38, 42, 56, 64, 68, 69 *and* 62 *kg. Do you agree with the claim that medicine B increases the weight significantly?*

Solution

$$\begin{aligned} n_1 &= 5, n_2 = 7 \\ \left.\begin{aligned} \bar{x} &= 46 \text{ kg} \\ \bar{y} &= 57 \text{ kg} \\ s_1 &= 7.62 \text{ kg} \\ s_2 &= 11.5 \text{ kg} \end{aligned}\right\} & \text{ From calculator} \end{aligned}$$

$$s = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(5)(7.62)^2 + (7)(11.5)^2}{5+7-2}} = 11.03 \text{ kg}$$

(i) Null Hypothesis $H_0 : \mu_1 = \mu_2$, i.e., there is no significant difference between the medicines A and B as regards their effect on the increase in weight.

(ii) Alternative Hypothesis $H_1 : \mu_1 > \mu_2$ (One tailed test)

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic $t = \dfrac{\bar{x} - \bar{y}}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} = \dfrac{46-57}{11.03\sqrt{\dfrac{1}{5} + \dfrac{1}{7}}} = -1.7$

$$|t| = 1.7$$

(v) Critical value: $v = n_1 + n_2 - 2 = 5 + 7 - 2 = 10$

$$t_{0.05}\,(v = 10) = 1.812.$$

(vi) Decision: Since $|t| < t_{0.05}$, the null hypothesis is accepted at 5% level of significance, i.e., the medicines *A* and *B* do not differ significantly as regards their effect on increase in weight.

Example 6

The following data represent the biological values of protein from cow's milk and buffalo's milk at a certain level:

Cow's milk	1.82	2.02	1.88	1.61	1.81	1.54
Buffalo's milk	2.00	1.83	1.86	2.03	2.19	1.88

Examine if the average values of protein in the two samples significantly differ.

Solution

Here, $n_1 = n_2 = 6$ and two samples are independent.

$$\begin{aligned} n &= 6 \\ \left.\begin{aligned} \bar{x}_1 &= 1.78 \\ \bar{x}_2 &= 1.965 \\ s_1 &= 0.16 \\ s_2 &= 0.124 \end{aligned}\right\} & \text{ From calculator} \end{aligned}$$

(i) Null Hypothesis H_0: $\mu_1 = \mu_2$, i.e., there is no significant difference in average values of proteins in two milk samples.

(ii) Alternative Hypothesis H_1: $\mu_1 \neq \mu_2$ (Two tailed test)

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2 + s_2^2}{n-1}}} = \dfrac{1.78 - 1.965}{\sqrt{\dfrac{(0.16)^2 + (0.124)^2}{6-1}}} = -2.043$

$$|t| = 2.043$$

(v) Critical value: $v = 2n - 2 = 2(6) - 2 = 10$

$$t_{0.05}\,(v = 10) = 2.228$$

(vi) Decision: Since $|t| < t_{0.05}$, the null hypothesis is accepted at 5% level of significance, i.e., there is no significant difference in average values of proteins in two milk samples.

Example 7

A certain injection administered to 12-*patients resulted in the following changes of blood pressure:*

5, 2, 8, −1, 3, 0, 6, −2, 1, 5, 0, 4

Can it be concluded that the injection will be in general accompanied by an increase in blood pressure?

Solution

Here, 'the changes' $d = x - y$ in blood pressure are given, i.e., x is the final blood pressure after administering the injection and y is the initial blood pressure. It is required to test whether the mean blood pressure has increased, i.e., μ_1 is greater than μ_2.

$$n = 12, \ \sum d_i = 31, \ \sum d_i^2 = 185$$

$$\bar{d} = \frac{\sum d_i}{n} = \frac{31}{12} = 2.58$$

$$s = \sqrt{\frac{\sum d_i^2}{n} - \left(\frac{\sum d_i}{n}\right)^2} = \sqrt{\frac{185}{12} - \left(\frac{31}{12}\right)^2} = 2.96$$

(i) Null Hypothesis H_0: $\mu_1 = \mu_2$, i.e., mean blood pressure has not increased.
(ii) Alternative Hypothesis H_1: $\mu_1 > \mu_2$ (One tailed test)
(iii) Level of significance: $\alpha = 0.05$ (assumption)
(iv) Test statistic: $t = \dfrac{\bar{d}}{\left(\dfrac{s}{\sqrt{n-1}}\right)} = \dfrac{2.58}{\left(\dfrac{2.96}{\sqrt{12-1}}\right)} = 2.89$

$$|t| = 2.89$$

(v) Critical value: $\nu = n - 1 = 12 - 1 = 11$

$$t_{0.05}\,(\nu = 11) = 1.796$$

(vi) Decision: Since $|t| > t_{0.05}$, the null hypothesis is rejected, i.e., injection will in general accompanied by an increase in blood pressure.

Example 8

Scores obtained in a shooting competition by 10 *soldiers before and after intensive training are given below:*

Score before training	67	24	57	55	63	54	56	68	33	43
Score after training	70	38	58	58	56	67	68	75	42	38

Test whether the intensive training is useful at 0.05 *level of significance.*

Solution

Since both the scores belongs to same set of soldiers, scores can be regarded as correlated and no longer independent. Paired t-test is applied to check hypothesis.

$$n_1 = n_2 = n = 10$$

Calculation of paired-t

x	67	24	57	55	63	54	56	68	33	43
y	70	38	58	58	56	67	68	75	42	38
$d = x - y$	−3	−14	−1	−3	7	−13	−12	−7	−9	5
d^2	9	196	1	9	49	169	144	49	81	25

$$\sum d_i = -50, \quad \sum d_i^2 = 732$$

$$\bar{d} = \frac{\sum d_i}{n} = \frac{-50}{10} = -5$$

$$s = \sqrt{\frac{\sum d_i^2}{n} - \left(\frac{\sum d_i}{n}\right)^2} = \sqrt{\frac{732}{10} - \left(\frac{-50}{10}\right)^2} = 6.94$$

(i) Null Hypothesis H_0: $\mu_1 = \mu_2$, i.e., there is no significant effect of intensive training.
(ii) Alternative Hypothesis H_1: $\mu_1 < \mu_2$ (One tailed test)
(iii) Level of significance: $\alpha = 0.05$
(iv) Test statistic: $t = \dfrac{\bar{d}}{\left(\dfrac{s}{\sqrt{n-1}}\right)} = \dfrac{-5}{\left(\dfrac{6.94}{\sqrt{10-1}}\right)} = -2.16$

$$|t| = 2.16$$

(v) Critical value: $\nu = n - 1 = 10 - 1 = 9$

$$t_{0.05}\,(\nu = 9) = 1.96$$

(vi) Decision: Since $|t| > t_{0.05}$, the null hypothesis is rejected at 5% level of significance, i.e., intensive training is useful.

6.14 *t*-TEST: TEST OF SIGNIFICANCE FOR CORRELATION COEFFICIENTS

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be n pairs of observations of a random sample from a bivariate normal population and let r be the observed correlation coefficient in the sample. It is required to test if this sample correlation coefficient is significant of any correlation in the population, i.e., whether the value of the population correlation coefficient ρ is zero and the observed value of r has arisen due to fluctuation of sampling. The student's t statistic is given by

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \text{ with } \nu = n - 2$$

Example 1

A random sample of 18 *pairs of observations from a bivariate normal population gives a correlation coefficient of* 0.3. *Is it likely that variables are uncorrelated in the population?*

Solution

$$n = 18, \qquad r = 0.3$$

(i) Null Hypothesis H_0: $\rho = 0$, i.e., the variables are uncorrelated.
(ii) Alternative Hypothesis H_1: $\rho \neq 0$ (Two tailed test)
(iii) Level of significance: $\alpha = 0.05$ (assumption)
(iv) Test statistic: $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \dfrac{0.3\sqrt{18-2}}{\sqrt{1-(0.3)^2}} = 1.26$

$$|t| = 1.26$$

(v) Critical value: $v = n - 2 = 18 - 2 = 16$

$$t_{0.05}(v = 16) = 2.12$$

(vi) Decision: Since $|t| < t_{0.05}$, the null hypothesis is accepted at 5% level of significance, i.e., the variables are uncorrelated in the population.

Example 2

A random sample of 10 *nations gives a correlation coefficient of* 0.5 *between literacy rate and political stability. Is the relationship significant?*

Solution

$$n = 10, \quad r = 0.5$$

(i) Null Hypothesis $H_0 : \rho = 0$, i.e., there is no relationship between literacy rate and political stability.

(ii) Alternative Hypothesis $H_1 : \rho \neq 0$ (Two tailed test)

(iii) Level of significance $\alpha = 0.5$ (assumption)

(iv) Test statistic: $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \dfrac{0.5\sqrt{10-2}}{\sqrt{1-(0.5)^2}} = 1.63$

$$|t| = 1.63$$

(v) Critical value: $v = n - 2 = 10 - 2 = 8$

$$t_{0.05}(v = 8) = 2.306$$

(vi) Decision: Since $|t| < t_{0.05}$, the null hypothesis is accepted at 5% level of significance i.e., there is no relationship between literacy rate and political stability.

Example 3

Find the least value of r in samples of 18 *pairs of observations from a bivariate normal population, which is significant at 5% level.*

Solution

The value of r for $n = 18$ will be significant at 5% level if

$$\left|\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\right| \geq t_{0.05}(v = 16)$$

$$\left|\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\right| \geq 2.12$$

Squaring both the sides and putting $n = 18$,

$$\frac{r^2(18-2)}{1-r^2} \geq 4.5$$

$$16r^2 \geq 4.5 - 4.5r^2$$

$$20.5r^2 \geq 4.5$$

$$r^2 \geq 0.22$$

$$|r| \geq 0.47$$

Hence, the least value of r is 0.47 (numerically).

EXERCISE 6.3

1. A sample of 26 bulbs gives a mean life of 990 hours with a SD of 20 hours. The manufacturer claims that the mean life of bulbs is 1000 hours. Is the sample not up to standard?

   [Ans.: The sample is not up to the standard]

2. The average breaking strength of the steel rods is specified to be 18.5 thousand pounds. To test this, sample of 14 rods were tested. The mean and SD obtained were 17.85 and 1.955 respectively. Is the result of experiment significant?

   [Ans.: The result of experiment is not significant]

3. A random sample of six steel beams has a mean compressive strength of 58392 psi (pounds per square inch) with a SD of 648 psi. Use this information and level of significance $\alpha = 0.05$ to test whether the true average compressive strength of the steel from which this sample came is 58000 psi. Assume normality.

   [Ans.: The average compressive strength of the steel beam is not equal to 58000 psi]

4. A sample of 155 members has a mean of 67 and SD of 52. Is this sample has been taken from a large population of mean 70?

   [Ans.: The sample has not been taken from the given population]

5. The heights of 10 males of a given locality are found to be 70, 67, 62, 68, 61, 68, 70, 64, 64, 66 inches. Is it reasonable to believe that the average height is greater than 64 inches? Test at 5% significance level assuming that for 9 degrees of freedom $t = 1.833$ at $\alpha = 0.05$.

   [Ans.: The average height is greater than 64 inches]

6. A random sample from a company's very extensive files shows that the orders for a certain kind of machinery were filled respectively in 10, 12, 19, 14, 15, 18, 11 and 13 days. Use the level of significance $\alpha = 0.01$ to test the claim that on the average such orders are filled in 10.5 days.

Choose the alternative hypothesis so that rejection of null hypothesis $\mu = 10.5$ days implies that it takes longer than indicated.

[**Ans.:** The orders on average are filled in more than 10.5 days]

7. Producer of gutkha claims that the nicotine content in his gutkha on the average is 1.83 mg. Can this claim be accepted if a random sample of 8 gutkha of this type have the nicotine contents of 2, 1.7, 2.1, 1.9, 2.2, 2.1, 2, 1.6 mg? Use a 0.05 level of significance.

[**Ans.:** The null hypothesis is accepted]

8. Two horses *A* and *B* were tested according to the time (in seconds) to run a particular track with the following results:

Horse A	28	30	32	33	33	29	34
Horse B	29	30	30	24	27	29	

Test whether the two horses have the same running capacity.

[**Ans.:** The two horses do not have the same running capacity]

9. To examine the hypothesis that the husbands are more intelligent than the wives, an investigator took a sample of 10 couples and administered them a test which measures the IQ. The results are as follows:

Husbands	117	105	97	105	123	109	86	78	103	107
Wives	106	98	87	104	116	95	90	69	108	85

Test the hypothesis with a reasonable test at the level of significance of 0.05.

[**Ans.:** These is no significant difference in IQs]

10. Two independent samples of 8 and 7 items respectively had the following values:

Sample I	11	11	13	11	15	9	12	14
Sample II	9	11	10	13	9	8	10	-

Is the difference between the means of samples significant?

[**Ans.:** The difference between the mean of samples is not significant]

11. Random samples of specimens of coal from two mines A and B are drawn and their heat-producing capacity (in millions of calories/ton) were measured yielding the following results:

Mine A	8350	8070	8340	8130	8260	-
Mine B	7900	8140	7920	7840	7890	7950

Is there significant difference between the means of these two samples at 0.01 level of significance?

[**Ans.:** There is significant difference between the means of two samples]

12. A random sample of 27 pairs of observations from a bivariate normal population gives a correlation coefficient of 0.42. Is it likely that the variables are uncorrelated in the population?

[**Ans.:** correlated]

13. Find the least value of *r* in a sample of 27 pairs from a bivariate normal population which is significant at 5% level.

[**Ans.:** $|r| = 0.487$]

6.15 SNEDECOR'S *F*-TEST FOR RATIO OF VARIANCES

Let $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ be the values of two independent random samples of sizes n_1 and n_2 ($n_1 \le 30$, $n_2 \le 30$) with means $\overline{x}$ and $\overline{y}$ drawn from the normal population with mean μ and standard deviation σ. The test statistic of Snedecor's *F*-test in terms of unbiased estimates of standard deviations S_1 and S_2 of population is given by

$$F = \frac{S_1^2}{S_2^2}, \quad \text{where } S_1^2 > S_2^2$$

and

$$S_1^2 = \frac{\sum (x-\overline{x})^2}{n_1 - 1}$$

$$S_2^2 = \frac{\sum (y-\overline{y})^2}{n_2 - 1}$$

with numerator degree of freedom $v_1 = n - 1$ and denominator degree of freedom $v_2 = n_2 - 1$.

If s_1 and s_2 are standard deviations of samples then

$$s_1^2 = \frac{\sum (x-\overline{x})^2}{n_1}$$

$$s_2^2 = \frac{\sum (y-\overline{y})^2}{n_2}$$

$$\therefore \quad \sum (x-\overline{x})^2 = n_1 s_1^2$$

$$\sum (y-\overline{y})^2 = n_2 s_2^2$$

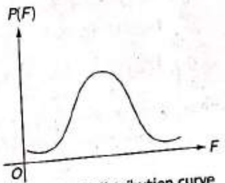


Fig. 6.6 *F*-distribution curve

Substituting in S_1^2 and S_2^2,

$$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1}$$

$$S_2^2 = \frac{n_2 s_2^2}{n_2 - 1}$$

The Snedecor's *F*-distribution is defined by

$$P(F) = cF^{\left(\frac{\nu_1-2}{2}\right)}\left(1+\frac{\nu_1}{\nu_2}F\right)^{-\left(\frac{\nu_1+\nu_2}{2}\right)}$$

where the constant *c* depends on ν_1 and ν_2. It is so chosen that the area under the curve is unity.

6.15.1 Properties of *F*-distribution

(i) *F*-distribution curve lies entirely in the first quadrant and is unimodal.
(ii) *F*-distribution is independent of the population variance σ^2 and depends on ν_1 and ν_2 only.
(iii) The mode of *F*-distribution is less than unity.
(iv) $F_{1-\alpha}(\nu_1, \nu_2) = \dfrac{1}{F_\alpha(\nu_2, \nu_1)}$

where $F_\alpha(\nu_2, \nu_1)$ is the value of *F* with ν_2 and ν_1 degrees of freedom such that the area under the *F*-distribution curve right of F_α is α.
(v) *F*-test is one tailed test (right tailed test).

6.15.2 Test of Significance for Ratio of Variances

Significant test is performed by means of Snedecer's *F*-table which provides 5% and 1% of points of significance for *F*. 5% points of *F* means that the area under the *F*-curve, to the right of the ordinate at a value of *F*, is 0.05. Further, *F*-table gives only single tail test. *F*-distribution is very useful for testing the equality of population means by comparing sample variances.

Working Rule

(i) Null Hypothesis $H_0: \sigma_1^2 = \sigma_2^2$
(ii) Alternative Hypothesis $H_1: \sigma_1^2 > \sigma_2^2$
(iii) Level of significance: Select the level of significance
(iv) Test statistic: $F = \dfrac{S_1^2}{S_2^2}$ where $S_1^2 > S_2^2$
(v) Critical value: Find the critical value (tabulated value) F_α at the given level of significance at degree of freedoms,

$$\nu_1 = n_1 - 1$$
$$\nu_2 = n_2 - 1$$

(vi) Decision: If $F < F_\alpha$ at the level of significance α, the null hypothesis is accepted. If $F > F_\alpha$ at the level of significance α, the null hypothesis is rejected.

Example 1

In two independent samples of sizes 8 and 10, the sum of squares of deviations of the sample values from the respective means were 84.4 and 102.6. Test whether the difference of variances of the population is significant or not. Use a 0.05 level of significance.

Solution

$$n_1 = 8, \quad n_2 = 10$$
$$\Sigma(x-\bar{x})^2 = 84.4, \quad \Sigma(y-\bar{y})^2 = 102.6,$$
$$S_1^2 = \frac{\Sigma(x-\bar{x})^2}{n_1-1} = \frac{84.4}{8-1} = 12.057$$
$$S_2^2 = \frac{\Sigma(y-\bar{y})^2}{n_2-1} = \frac{102.6}{10-1} = 11.4$$

(i) Null Hypothesis $H_0: \sigma_1^2 = \sigma_2^2$, i.e., the variances of two populations are equal.
(ii) Alternative Hypothesis $H_1: \sigma_1^2 > \sigma_2^2$
(iii) Level of significance: $\alpha = 0.05$
(iv) Test statistic: $F = \dfrac{S_1^2}{S_2^2} = \dfrac{12.057}{11.4} = 1.057$
(v) Critical value: $\nu_1 = n_1 - 1 = 8 - 1 = 7$

$$\nu_2 = n_2 - 1 = 10 - 1 = 9$$
$$F_{0.05}(\nu_1 = 7, \nu_2 = 9) = 3.29$$

(vi) Decision: Since $F < F_{0.05}$, the null hypothesis is accepted at 0.05 level of significance, i.e., there is no significant difference in variances of the population.

Example 2

The standard deviations calculated from two random samples of sizes 9 and 13 are 2.1 and 1.8 respectively. Can the samples be regarded as drawn from normal populations with the same SD?

Solution

$$n_1 = 9, \; n_2 = 13, \quad s_1 = 2.1, \quad s_2 = 1.8$$
$$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{9(2.1)^2}{9-1} = 4.96$$
$$S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{13(1.8)^2}{13-1} = 3.51$$

(i) Null Hypothesis H_0: $\sigma_1^2 = \sigma_2^2$, i.e., variances of two populations are equal.

(ii) Alternative Hypothesis H_1: $\sigma_1^2 > \sigma_2^2$

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: $F = \frac{S_1^2}{S_2^2} = \frac{4.96}{3.51} = 1.41$

(v) Critical value: $v_1 = n_1 - 1 = 9 - 1 = 8$

$$v_2 = n_2 - 1 = 13 - 1 = 12$$

$$F_{0.05}(v_1 = 8,\ v_2 = 12) = 2.85$$

(vi) Decision: Since $F < F_{0.05}$, the null hypothesis is accepted at 5% level of significance, i.e., the samples can be regarded as drawn from normal population with same SD.

Example 3

Two random samples are drawn from two populations and the following results were obtained:

Sample I	16	17	18	19	20	21	22	24	26	27		
Sample II	19	22	25	25	26	28	29	30	31	32	35	36

Find the variances of the two samples and test whether the two populations have the same variances.

Solution

$$n_1 = 10, \quad n_2 = 12$$

$$\left.\begin{aligned} \bar{x}_1 &= 21 \\ \bar{x}_2 &= 28 \\ s_1 &= 3.55 \\ s_2 &= 4.98 \end{aligned}\right\} \text{From calculator}$$

$$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{10(3.55)^2}{10 - 1} = 14$$

$$S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{12(4.98)^2}{12 - 1} = 27.05$$

(i) Null Hypothesis H_0: $\sigma_1^2 = \sigma_2^2$, i.e., two populations have the same variances.

(ii) Alternative Hypothesis H_1: $\sigma_1^2 > \sigma_2^2$

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: Since $S_2^2 > S_1^2$,

$$F = \frac{S_2^2}{S_1^2} = \frac{27.05}{14} = 1.93$$

(v) Critical value: $v_1 = n_1 - 1 = 10 - 1 = 9$

$$v_2 = n_2 - 1 = 12 - 1 = 11$$

$$F_{0.05}(v_2 = 11,\ v_1 = 9) = 3.10$$

(vi) Decision: Since $F < F_{0.05}$, the null hypothesis is accepted at 5% level of significance, i.e., two populations have the same variances.

Example 4

In a test given to two groups of students drawn from two normal populations, the marks obtained were as follows:

Group I	18	20	36	50	49	36	34	49	41
Group II	29	28	26	35	30	44	46		

Examine at 5% level, whether the two populations have the same variances.

Solution

$$n_A = 9$$

$$n_B = 7$$

$$\left.\begin{aligned} \bar{x} &= 37 \\ \bar{y} &= 34 \\ s_1 &= 11.225 \\ s_2 &= 7.426 \end{aligned}\right\} \text{From calculator}$$

$$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{9(11.2225)^2}{9 - 1} = 141.75$$

$$S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{7(7.426)^2}{7 - 1} = 64.3$$

(i) Null Hypothesis $H_0: \sigma_1^2 = \sigma_2^2$, i.e., the two populations have same variances.

(ii) Alternative Hypothesis $H_1: \sigma_1^2 \neq \sigma_2^2$

(iii) Level of significance: $\alpha = 0.05$

(iv) Test statistic: $F = \dfrac{S_1^2}{S_2^2} = \dfrac{141.75}{64.33} = 2.203$

(v) Critical value: $v_1 = n_1 - 1 = 9 - 1 = 8$

$v_2 = n_2 - 1 = 7 - 1 = 6$

$F_{0.05}(v_1 = 8, v_2 = 6) = 4.15$

(vi) Decision: Since $F < F_{0.05}$, the null hypothesis is accepted at 5% level of significance, i.e., the two populations have the same variances.

Example 5

A group of 10 rats fed on diet A and another group of 8 rats fed on diet B recorded following increase in weight:

Diet A	5	6	8	1	12	4	3	9	6	10	gm
Diet B	2	3	6	8	1	10	2	8			gm

Find, if the variances are significantly different?

Solution

$n_1 = 10, \quad n_2 = 8$

$\left.\begin{array}{l} s_1 = 3.2 \\ s_2 = 3.23 \end{array}\right\}$ From calculator

$$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{10(3.2)^2}{10-1} = 11.38$$

$$S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{8(3.23)^2}{8-1} = 11.92$$

(i) Null Hypothesis H_0: $\sigma_1^2 = \sigma_2^2$, i.e., the there is no significant difference in variances.

(ii) Alternative Hypothesis H_1: $\sigma_1^2 > \sigma_2^2$

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: Since $S_2^2 > S_1^2$,

$$F = \frac{S_2^2}{S_1^2} = \frac{11.92}{11.33} = 1.05$$

(v) Critical value: $v_1 = n_1 - 1 = 10 - 1 = 9$

$v_2 = n_2 - 1 = 8 - 1 = 7$

$F_{0.05}(v_2 = 7, v_1 = 9) = 3.29$

(vi) Decision: Since $F < F_{0.05}$, the null hypothesis is accepted at 5% level of significance, i.e., the two variances are not significantly different.

Example 6

Two random samples gave the following data:

	Size	*Mean*	*Variance*
Sample I	8	9.6	1.2
Sample II	11	16.5	2.5

Can we conclude that the two samples have been drawn from the same normal population?

Solution

A normal distribution has two parameters, mean μ and variance σ^2. To conclude that the two samples have been drawn from the same normal population, we have to test for

(i) Equality of two means H_0 $(\mu_1 = \mu_2)$ by *t*-test

(ii) Equality of two variances H_0 $(\sigma_1^2 = \sigma_2^2)$ by *F*-test.

F-test:

$n_1 = 8, \quad n_2 = 11, \quad \bar{x}_1 = 9.6, \quad \bar{x}_2 = 16.5, \quad s_1^2 = 1.2, \quad s_2^2 = 2.5$

$$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{8(1.2)}{8-1} = 1.37$$

$$S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{11(2.5)}{11-1} = 2.75$$

(i) Null Hypothesis H_0: $\sigma_1^2 = \sigma_2^2$, i.e., variances of two populations are equal.

(ii) Alternative Hypothesis H_1: $\sigma_1^2 > \sigma_2^2$

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: Since $S_2^2 > S_1^2$,

$$F = \frac{S_2^2}{S_1^2} = \frac{2.75}{1.37} = 2.007$$

(v) Critical value: $v_1 = n_1 - 1 = 8 - 1 = 7$

$v_2 = n_2 - 1 = 11 - 1 = 10$

$F_{0.05}(v_2 = 10, v_1 = 7) = 3.64$

(vi) Decision: Since $F < F_{0.05}$, the null hypothesis is accepted at 5% level of significance, i.e., two populations have the same variances.

t-test:

$$s=\sqrt{\frac{n_1s_1^2+n_2s_2^2}{n_1+n_2-2}}=\sqrt{\frac{8(1.2)+11(2.5)}{8+11-2}}=1.48$$

(i) Null Hypothesis H_0: $\mu_1=\mu_2$, i.e., means of two populations are equal.

(ii) Alternative Hypothesis H_1: $\mu_1\neq\mu_2$ (Two tailed test)

(iii) Level of significance: $\alpha=0.05$ (assumption)

(iv) Test statistic: $t=\dfrac{\bar{x}_1-\bar{x}_2}{s\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}=\dfrac{9.6-16.5}{1.48\sqrt{\dfrac{1}{8}+\dfrac{1}{11}}}=-10.03$

$$|t|=10.03$$

(v) Critical value: $\nu_1=n_1+n_2-2=8+11-2=17$

$$t_{0.05}(\nu=17)=2.11$$

(vi) Decision: Since $|t|>t_{0.05}$, the null hypothesis is rejected at 5% level of significance, i.e., two populations have not same means.

Hence, the two samples could not have been drawn from the same normal population.

Example 7

Two nicotine contents in two random samples of tobacco are given below:

Sample I	21	24	25	26	27	
Sample II	22	27	28	30	31	36

Can we say that two samples came from the same population?

Solution

F-test:

$$n_1=5,\quad n_2=6$$

$$\left.\begin{aligned}\bar{x}_1&=24.6\\ \bar{x}_2&=29\\ s_1&=2.06\\ s_2&=4.24\end{aligned}\right\}\text{From calculator}$$

$$S_1^2=\frac{n_1s_1^2}{n_1-1}=\frac{5(2.06)^2}{5-1}=5.30$$

$$S_2^2=\frac{n_2s_2^2}{n_2-1}=\frac{6(4.24)^2}{6-1}=21.57$$

(i) Null Hypothesis H_0: $\sigma_1^2=\sigma_2^2$, i.e., variances of two populations are equal.

(ii) Alternative Hypothesis H_1: $\sigma_1^2>\sigma_2^2$

(iii) Level of significance: $\alpha=0.05$ (assumption)

(iv) Test statistic: Since $S_2^2>S_1^2$,

$$F=\frac{S_2^2}{S_1^2}=\frac{21.57}{5.30}=4.07$$

(v) Critical value: $\nu_1=n_1-1=5-1=4$

$$\nu_2=n_2-1=6-1=5$$

$$F_{0.05}(\nu_2=5,\ \nu_1=4)=6.26$$

(vi) Decision: Since $F<F_{0.05}$, the null hypothesis is accepted at 5% level of significance, i.e., the two populations have the same variances.

t-test:

$$s=\sqrt{\frac{n_1s_1^2+n_2s_2^2}{n_1+n_2-2}}=\sqrt{\frac{5(2.06)^2+6(4.24)^2}{5+6-2}}=14.34$$

(i) Null Hypothesis H_0: $\mu_1=\mu_2$, i.e., means of two populations are equal.

(ii) Alternative Hypothesis H_1: $\mu_1\neq\mu_2$ (Two tailed test)

(iii) Level of significance: $\alpha=0.05$ (assumption)

(iv) Test statistic: $t=\dfrac{\bar{x}_1-\bar{x}_2}{s\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}=\dfrac{24.6-29}{14.34\sqrt{\dfrac{1}{5}+\dfrac{1}{6}}}=-0.51$

$$|t|=0.51$$

(v) Critical value: $\nu=n_1+n_2-2=5+6-2=9$

$$t_{0.05}(\nu=9)=2.262$$

(vi) Decision: Since $|t|<t_{0.05}$, the null hypothesis is accepted at 5% level of significance, i.e., two populations have same means.

Hence, two samples came from the same population.

EXERCISE 6.4

1. If two independent samples of sizes $n_1=13$ and $n_2=7$ are taken from a normal population. What is the probability that the variance of the first sample will be at least four times as large as that of the second sample?

[Ans.: 0.05]

2. The standard deviations calculated from two random samples of size 9 and 13 are 2 and 1.9 respectively. Can the samples be regarded as drawn from the normal populations with the same standard deviation?
[Ans.: The samples can be regarded as drawn from the normal populations with the same standard deviation]

3. Two samples are drawn from two normal populations. From the following data test whether the two samples have the same variance at 5% level?

Sample I	60	65	71	74	76	82	85	87		
Sample II	61	66	67	85	78	63	85	86	88	91

[Ans.: Two samples have the same variances]

4. The time taken by workers in performing a job by method I and method II is given below.

Method I	20	16	26	27	22		
Method II	27	33	42	35	32	34	38

Do the data show that the variances of time distribution in a population from which these samples are drawn do not differ significantly?
[Ans.: The variances of time distribution in a population from which the samples are drawn do not differ significantly]

5. Following results were obtained from two samples, each drawn from two different population A and B:

Population	A	B
Sample	I	II
Sample size	25	17
Sample SD	3	2

Test the hypothesis that the variance of brand *A* is more than that of *B*.
[Ans.: Variance of brand *A* is not more than the variance of brand *B*]

6. In a laboratory experiment two samples gave the following results:

Sample	*Size*	*Sample mean*	*Sum of squares of deviation from the mean*
1	10	15	90
2	12	14	108

Test the equality of sample variances at 5% level of significance.
[Ans.: The two population have the same variances]

6.16 CHI-SQUARE (χ^2) TEST

The chi-square (χ^2) test is a useful measure of comparing experimentally obtained results with those expected theoretically and based on hypothesis. It is used as a test statistic in testing a hypothesis that provides a set of theoretical frequencies with which observed frequencies are compared. The magnitude of discrepancy between observed and theoretical frequencies is given by the quantity χ^2 (pronounced as chi-square). If $\chi^2=0$, the observed and expected frequencies completely coincide. As the value of χ^2 increases, the discrepancy between the observed and theoretical frequency decreases. If $f_{o_1}, f_{o_2}, \dots, f_{o_n}$ be a set of observed frequencies and $f_{e_1}, f_{e_2}, \dots, f_{e_n}$ be the corresponding set of expected (or theoretical) frequencies then χ^2 is defined by

$$\chi^2=\frac{(f_{o_1}-f_{e_1})^2}{f_{e_1}}+\frac{(f_{o_2}-f_{e_2})^2}{f_{e_2}}+\dots+\frac{(f_{o_n}-f_{e_n})^2}{f_{e_n}}=\sum\frac{(f_o-f_e)^2}{f_e}$$

with $n-1$ degrees of freedom.

Note

If the data is given in a series of n numbers then degrees of freedom $v=n-1$
In case of binomial distribution, $v=n-1$
In case of Poisson distribution, $v=n-2$
In case of normal distribution, $v=n-3$

6.16.1 Chi-Square Distribution

If $x_1, x_2, \dots, x_n$ are n independent normal variates with mean zero and standard deviation unity then $x_1^2+x_2^2+\dots+x_n^2$ is a random variate having χ^2 distribution with probability density function given by

$$P(\chi^2)=y_0(\chi^2)^{\frac{v-1}{2}}e^{-\frac{\chi^2}{2}}$$

where $v=$ degrees of freedom $=n-1$ and $y_0=$ constant depending on the degrees of freedom

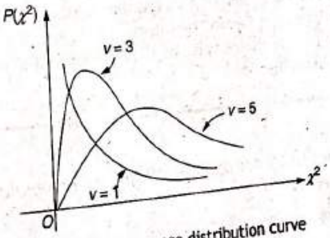


Fig. 6.7 Chi-square distribution curve

6.16.2 Properties of χ^2-Distribution

(i) Chi-Square test is always positively skewed.
(ii) The mean of chi-square distribution is the number of degrees of freedom.
(iii) The standard deviation of chi-square distribution = $\sqrt{2v}$.
(iv) Chi-square values increases with the increase in degrees of freedom.
(v) The value of χ^2 lies between zero and infinity.
(vi) For different values of degrees of freedom, the shape of the curve will be different.

6.17 CHI-SQUARE TEST: GOODNESS OF FIT

The values of χ^2 is used to test whether the deviations of the observed frequencies from the expected frequencies are significant or not. It is also used to fit a set of observations to a given distribution. Hence, chi-square test provides a test of goodness of fit and may be used to examine the validity of some hypothesis about an observed frequency distribution.

Test of Significance

Let $f_{o_1}, f_{o_2}, \ldots, f_{o_n}$ be a set of observed frequencies and $f_{e_1}, f_{e_2}, \ldots, f_{e_n}$ be the corresponding set of expected or theoretical frequencies. The χ^2 statistic is given by

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Working Rule

(i) Set up a null hypothesis.
(ii) Set up an alternative hypothesis.
(iii) Set a level of significance α.
(iv) Calculate χ^2.
(v) Find the degree of freedom and find the corresponding value of χ^2 at given level of significance α.
(vi) If the calculated value of χ^2 is less than tabulated value of χ^2 at the level of significance α, the null hypothesis is accepted. If calculated value of χ^2 is more than tabulated value of χ^2 at the level of significance α, the null hypothesis is rejected.

Example 1

A dice was thrown 132 *times and the following frequencies were observed:*

No. obtained	1	2	3	4	5	6	Total
Frequency	15	20	25	15	29	28	132

Test the hypothesis that the dice is unbiased.

Solution

$n = 6$

(i) Null Hypothesis H_0: The dice is unbiased.
(ii) Alternative Hypothesis H_1: The dice is biased.
(iii) Level of significance: $\alpha = 0.05$ (assumption)
(iv) Test statistic:

Expected frequency of each number $f_e = \frac{132}{6} = 22$

No. obtained	*Observed frequency,* f_o	*Expected frequency,* f_e	$\frac{(f_o - f_e)^2}{f_e}$
1	15	22	2.23
2	20	22	0.18
3	25	22	0.41
4	15	22	2.23
5	29	22	2.23
6	28	22	1.64
			$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 8.92$

(v) Critical value: $v = n - 1 = 6 - 1 = 5$

$$\chi^2_{0.05}(v = 5) = 11.07$$

(vi) Decision: Since $\chi^2 < \chi^2_{0.05}$, the null hypothesis is accepted at 5% level of significance, i.e., the dice is unbiased.

Example 2

The number of car accidents in a metropolitan city was found to be 20, 17, 12, 6, 7, 15, 8, 5, 16 *and* 14 *per month respectively. Use χ^2 test to check whether these frequencies are in agreement with the belief that the occurrence of accidents was the same during* 10 *months period. Test at* 5% *level of significance.*

Solution

$n = 10$

(i) Null Hypothesis H_0: Occurrence of accident was same during 10 months period.

(ii) Alternative Hypothesis H_1: Occurrence of accidents was not same during 10 months period.

(iii) Level of significance: $\alpha = 0.05$

(iv) Test statistic: If occurrence of accidents is same, the expected frequency of accidents per month

$$f_e = \frac{20+17+12+6+7+15+8+5+16+14}{10} = 12$$

Observed frequency, f_o	*Expected frequency, f_e*	$\frac{(f_o - f_e)^2}{f_e}$
20	12	5.33
17	12	2.08
12	12	0
6	12	3
7	12	2.08
15	12	0.75
8	12	1.33
5	12	4.08
16	12	1.33
14	12	0.33
		$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 20.31$

(v) Critical value: $v = n - 1 = 10 - 1 = 9$

$$\chi^2_{0.05} \ (v = 9) = 16.92$$

(vi) Decision: Since $\chi^2 > \chi^2_{0.05}$, the null hypothesis is rejected at 5% level of significance, i.e., occurrence of accidents was not same during 10 months period.

Example 3

200 digits were chosen at random from a set of tables, The frequency of the digits are shown below:

Digits	0	1	2	3	4	5	6	7	8	9
Frequency	18	19	23	21	16	25	22	20	21	15

Use the χ^2-test to access the correctness of the hypothesis that the digits were distributed in equal number in the tables from which these were chosen.

Solution

$$n = 10$$

(i) Null Hypothesis H_0: The digits were distributed in equal number in the tables.

(ii) Alternative Hypothesis H_1: The digits were not distributed in equal number in the tables.

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: Expected frequency of each digit $f_e = \frac{200}{10} = 20$

Observed frequency, f_o	*Expected frequency, f_e*	$\frac{(f_o - f_e)^2}{f_e}$
18	20	0.2
19	20	0.05
23	20	0.45
21	20	0.05
16	20	0.8
25	20	1.25
22	20	0.2
20	20	0
21	20	0.05
15	20	1.25
		$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 4.3$

(v) Critical value: $v = n - 1 = 10 - 1 = 9$

$$\chi^2_{0.05} \ (v = 9) = 16.92$$

(vi) Decision: Since $\chi^2 < \chi^2_{0.05}$, the null hypothesis is accepted at 5% level of significance, i.e., the digits were distributed in equal number in the table.

Example 4

Theory predicts that the proportion of beans in the four groups A, B, C, D should be 9 : 3 : 3: 1. In an experiment among 1600 beans, the numbers in the four groups were 882, 313, 287 and 118. Does the experimental results support the theory?

Solution

$$n = 4$$

(i) Null Hypothesis H_0: The proportion of the beans in the four groups A, B, C, D is $9:3:3:1$.

(ii) Alternative Hypothesis H_1: The proportion of the beans in the four groups A, B, C, D is not $9:3:3:1$.

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic:

Group	*Observed Frequency, f_o*	*Expected frequency, f_e*	$\frac{(f_o - f_e)^2}{f_e}$
A	882	$\frac{9}{16} \times 1600 = 900$	0.36
B	313	$\frac{3}{16} \times 1600 = 300$	0.56
C	287	$\frac{3}{16} \times 1600 = 300$	0.56
D	118	$\frac{1}{16} \times 1600 = 100$	3.24
			$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 4.72$

(v) Critical value: $\nu = n - 1 = 4 - 1 = 3$

$$\chi^2_{0.05} \ (\nu = 3) = 7.81$$

(vi) Decision: Since $\chi^2 < \chi^2_{0.05}$ the null hypothesis is accepted at 5% level of significance, i.e., experimental results support the theory and the proportion of the beans is $9:3:3:1$.

Example 5

The following mistakes per page were observed in a book:

No. of mistakes per page	0	1	2	3	4
No. of pages	211	90	19	5	0

Fit a Poisson distribution and test the goodness of fit.

Solution

(i) Null Hypothesis H_0: The mistakes follow Poisson distribution and Poisson distribution can be fitted to the data.

(ii) Alternative Hypothesis H_1: The mistakes do not follow Poisson distribution.

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: The expected frequencies by Poisson distribution are given by

$$\text{Expected frequency } f_e = Np = N\left(\frac{e^{-\lambda}\lambda^x}{x!}\right), x = 0,1,2,3,4$$

$$\lambda = \frac{\Sigma fx}{N} = \frac{211(0) + 90(1) + 19(2) + 5(3) + 0(4)}{211 + 90 + 19 + 5 + 0} = 0.44$$

$$f_e = Np = 325\left(\frac{e^{-0.44} 0.44^x}{x!}\right), \quad x = 0,1,2,3,4$$

Expected or Theoretical frequency

x	0	1	2	3	4
f_e	209.31	92.1	20.26	2.97	0.33

When expected frequencies are less than 10, classes are grouped together.

No. of mistakes	*Observed frequency f_o*	*Expected frequency f_e*	$f_o - f_e$	$\frac{(f_o - f_e)^2}{f_e}$
0	211	209.31	1.69	0.014
1	90	92.10	–2.1	0.048
2	19	20.26	0.44	0.008
3	5	2.97		
4	0	0.33		
				$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 0.07$

(v) Critical value: The number of degrees of freedom is 1 for each class. There are 5 classes originally. Hence, the degrees of freedom originally is 5. Since the classes are reduced by 2, the degrees of freedom is reduced by 2. Further, while calculating the parameter λ, two sums Σfx and Σf are used. Hence, the degrees of freedom is again reduced by 2.

Hence, the number of degrees of freedom $\gamma = 5 - (2 + 2) = 1$

$$\chi^2_{0.05} = 3.84$$

(vi) Decision: Since $\chi^2 < \chi^2_{0.05}$, the null hypothesis is accepted at 5% level of significance, i.e., the mistakes follow Poisson's distribution.

Example 6

A set of five similar coins is tossed 320 times and result is obtained as follows:

No. of heads	0	1	2	3	4	5
Frequency	6	27	72	112	71	32

Test the hypothesis that the data follow a binomial distribution.

Solution

(i) Null Hypothesis H_0: The data follow a binomial distribution.
(ii) Alternative Hypothesis H_1: The data do not follow binomial distribution.
(iii) Level of significance: $\alpha = 0.05$
(iv) Test statistic: Probability of getting a head $p = \frac{1}{2}$

Probability of getting a tail $q = \frac{1}{2}$
By binomial distribution,

$$p(x) = {}^nC_x p^x q^{n-x} = {}^5C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{5-x}, \quad x = 0,1,2,3,4,5$$

$$N = 320$$

$$\text{Expected frequency } f_e = Np(x) = 320\left[{}^5C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{5-x}\right], x = 0,1,2,3,4,5$$

Expected or theoretical frequency

x	0	1	2	3	4	5
f_e	10	50	100	100	50	10

No. of heads	Observed frequency f_o	Expected frequency f_e	$f_o - f_e$	$\frac{(f_o - f_e)^2}{f_e}$
0	6	10	−4	1.6
1	27	50	−23	10.58
2	72	100	−28	7.84
3	112	100	12	1.44
4	71	50	21	8.82
5	32	10	22	48.4
				$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 78.68$

(v) Critical value: $v = n - 1 = 6 - 1 = 5$

$$\chi^2_{0.05} = 11.07$$

(vi) Decision: Since $\chi^2 > \chi^2_{0.05}$ at 5% level of significance, the null hypothesis is rejected, i.e., the data do not follow the binomial distribution.

Example 7

Fit the equation of the best fitting normal curve to the following data:

x	135	145	155	165	175	185	195	205	Total
f	2	14	22	25	19	13	3	2	100

Compare the theoretical and observed frequencies. Using χ^2 test find goodness of fit. Given that $\mu = 165.6$ and $\sigma = 15.02$.

Solution

$$\mu = 165.6, \sigma = 15.02, N = \Sigma f = 100$$

The data is first converted into class intervals with inclusive series

Class interval	Lower class X	$Z = \frac{X - \mu}{\sigma}$	Area from 0 to Z	Area in class interval	Expected frequencies
130–140	130	−2.37	0.4911	0.0357	$3.57 \simeq 4$
140–150	140	−1.70	0.4554	0.1046	$10.46 \simeq 11$
150–160	150	−1.04	0.3508	0.2065	$20.65 \simeq 21$
160–170	160	−0.37	0.1443	0.2584	$25.84 \simeq 26$
170–180	170	0.29	0.1141	0.2174	$21.74 = 21$
180–190	180	0.96	0.3315	0.1159	$11.59 \simeq 12$
190–200	190	1.62	0.4474	0.0416	$4.16 \simeq 4$
200–210	200	2.29	0.4890	0.0095	$0.95 = 1$
210–220	210	2.96	0.4985		

Calculation of χ^2

When expected frequencies are less than 10, classes are grouped together.

x	Observed frequency f_o	Expected frequency f_e	$f_o - f_e$	$\frac{(f_o - f_e)^2}{f_e}$
135	2	4	1	0.067
145	14	11		
155	22	21	1	0.048
165	25	26	−1	0.038
175	19	21	−2	0.19
185	13	12	1	0.0588
195	3	4		
205	2	1		
				$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 0.4018$

Critical value: There are 5 frequencies. While calculating mean and standard deviation, three sums Σf, Σfx, and Σfx^2 are used. Hence, the number of degrees of freedom $v = 5 - 3 = 2$

$\chi^2_{0.05} = 5.99$

Since $\chi^2 < \chi^2_{0.05}$ at 5% level of significance, the fit is good and the distribution is nearly normal.

6.18 CHI-SQUARE TEST FOR INDEPENDENCE OF ATTRIBUTES

In statistics, sometimes we have to deal with attributes or qualitative characters, which cannot be measured accurately, although items can be divided into two or more categories w.r.t. the attributes. Let A and B be two attributes of the population. A can be divided into m categories $A_1, A_2 \ldots, A_m$ and B can be divided into n categories $B_1, B_2, \ldots, B_n$. The data can be shown in the form of a two-way table with m rows and n columns, as in a bivariate frequency distribution. This two-way frequency table for attributes is known as $m \times n$ contingency table. The frequency of observations belonging to both the categories A_i and B_i simultaneously is shown in the cell at the i-th row and j-th column and denoted by (A_iB_j). Similarly (A_i) and (B_j) denote the frequency of items belonging to categories A_i and B_j respectively and N, the total frequency.

(3 × 4) contingency table

		Attribute B				
		B_1	B_2	B_3	B_4	*Total*
Attribute A	A_1	$(A_1 B_1)$	$(A_1 B_2)$	$(A_1 B_3)$	$(A_1 B_4)$	(A_1)
	A_2	$(A_2 B_1)$	$(A_2 B_2)$	$(A_2 B_3)$	$(A_2 B_4)$	(A_2)
	A_3	$(A_3 B_1)$	$(A_3 B_2)$	$(A_3 B_3)$	$(A_3 B_4)$	(A_3)
	Total	(B_1)	(B_2)	(B_3)	(B_4)	N

Independence of Attributes

Two attributes A and B are said to be independent if they are not related to each other. If two attributes A and B are not independent, they are associated on the basis of cell frequencies. It is required to test whether two attributes A and B are associated or not. Under null hypothesis H_0 (attributes are independent), the expected frequency f_e of any cell is given by

$$f_e = \frac{(\text{Row total}) \times (\text{Column total})}{\text{Total frequency}} = \frac{(A_i)(B_j)}{N}$$

Then test statistic is given by

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

with degree of freedom v = (number of row − 1) (number of columns − 1)

If the calculated value of χ^2 is less than tabulated value of χ^2 at the given level of significance α for degree of freedom v, the null hypothesis is accepted and attributes are said to be independent. If calculated value of χ^2 is more than tabulated value of χ^2 at given level of significance α for degree of freedom v, the null hypothesis is rejected.

Yate's Correction

In a 2 × 2 table, there is only one degree of freedom. If any of the expected frequency is less than 10, Yate's correction is applied in chi-square formula.

$$\chi^2 = \sum \left[\frac{\{|f_o - f_e| - 0.5\}^2}{f_e} \right]$$

Example 1

A total of 3759 individual were interviewed in a public opinion survey on a political proposal. Of them 1872 were men and the rest were women. A total of 2257 individuals were in favour of the proposal and 917 were opposed to it. A total of 243 men were undecided and 442 women were opposed to it. Do you justify or contradict the hypothesis that there is no association between sex and attitude at 5% level of significance?

Solution

$N = 3759$

Opinion about political proposal

	Favoured	*Opposed*	*Undecided*	*Total*
Men	1154	475	243	1872
Women	1103	442	342	1887
Total	2257	917	585	3759

(i) Null Hypothesis H_0: There is no association between sex and attitude i.e., sex and attitude are independent.

(ii) Alternative Hypothesis H_1: There is association between sex and attitude.

(iii) Level of significance: $\alpha = 0.05$

(iv) Test statistic:

Calculation of χ^2

Observed Frequency f_o	*Expected Frequency* $f_e = \frac{(A_i)(B_j)}{N}$	$\frac{(f_o - f_e)^2}{f_e}$
1154	$\frac{1872 \times 2257}{3759} \approx 1124$	0.8
475	$\frac{1872 \times 917}{3759} \approx 457$	0.71
243	$\frac{1872 \times 585}{3759} \approx 291$	7.92
1103	$\frac{1887 \times 2257}{3759} \approx 1133$	0.79
442	$\frac{1887 \times 917}{3759} \approx 460$	0.70
342	$\frac{1887 \times 585}{3759} \approx 294$	7.84
		$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 18.76$

(v) Critical value: $v = (r-1)(c-1) = (2-1)(3-1) = 2$

$$\chi^2_{0.05}\ (v = 2) = 5.99$$

(vi) Decision: Since $\chi^2 > \chi^2_{0.05}$, the null hypothesis is rejected at 5% level of significance, i.e., there is association between sex and attitude.

Example 2

A sample of 400 students of undergraduate and 400 students of postgraduate classes was taken to know their opinion about autonomous colleges. 290 of the undergraduate and 310 of the postgraduate students favoured the autonomous status. Present these facts in the form of a table and test at 5% level of significance, that the opinion regarding autonomous status of colleges is independent of the level of classes of students.

Solution

$$N = 800$$

Opinion about autonomous colleges

	Favoured	*Not favoured*	*Total*
Undergraduate	290	110	400
Postgraduate	310	90	400
Total	600	200	800

(i) Null Hypothesis H_0: There is no relation between the classes of students and opinion, i.e., two attributes are independent.

(ii) Alternative Hypothesis H_1: There is relation between the classes of students and opinion.

(iii) Level of significance: $\alpha = 0.05$

(iv) Test statistic:

Observed Frequency f_o	*Expected frequency* $f_e = \frac{(A_i)(B_j)}{N}$	$\frac{(f_o - f_e)^2}{f_e}$
290	$\frac{400 \times 600}{800} = 300$	0.33
110	$\frac{400 \times 200}{800} = 100$	1.00
310	$\frac{400 \times 600}{800} = 300$	0.33
90	$\frac{400 \times 200}{800} = 100$	1.00
		$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 2.66$

(v) Critical value: $v = (r-1)(c-1) = (2-1)(2-1) = 1$

$$\chi^2_{0.05}\ (v - 1) = 3.81$$

(vi) Decision: Since $\chi^2 < \chi^2_{0.05}$, the null hypothesis is accepted at 5% level of significance, i.e., there is no relation between the classed of students are opinion.

Example 3

In an experiment on immunisation of cattle from tuberculosis the following results were obtained:

	Affected	Not affected	Total
Inoculated	267	27	294
Not inoculated	757	155	912
Total	1024	182	1206

Use χ^2-test to determine the efficiency of vaccine in preventing the tuberculosis.

Solution

$N = 1206$

(i) Null Hypothesis H_0: There is no relation between inoculation and effect on disease, i.e., two attributes are independent.

(ii) Alternative Hypothesis H_1: There is relation between inoculation and effect on disease.

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic:

Observed Frequency f_o	Expected frequency $f_e = \frac{(A_i)(B_j)}{N}$	$\frac{(f_o - f_e)^2}{f_e}$
267	$\frac{294 \times 1024}{1206} \approx 250$	1.156
27	$\frac{294 \times 182}{1206} \approx 44$	6.568
757	$\frac{912 \times 1024}{1206} \approx 774$	0.37
155	$\frac{912 \times 182}{1206} \approx 138$	2.09
		$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 10.19$

(v) Critical value: $v = (r - 1)(c - 1) = (2 - 1)(2 - 1) = 1$

$$\chi^2_{0.05}(v = 1) = 3.84$$

(vi) Decision: Since $\chi^2 > \chi^2_{0.05}$, the null hypothesis is rejected at 5% level of significance, i.e., vaccine is effective in preventing tuberculosis.

Example 4

Given the following contingency table for hair colour and eye colour. Find the value of χ^2. Is there good association between the two?

Eye colour	Hair colour			Total
	Fair	Brown	Black	
Blue	15	5	20	40
Grey	20	10	20	50
Brown	25	15	20	60
Total	60	30	60	150

Solution

$N = 150$

(i) Null Hypothesis H_0: There is no association between two attributes, hair and eye colours.

(ii) Alternative Hypothesis H_1: There is association between two attributes, hair and eye colours.

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic:

Observed Frequency f_o	Expected frequency $f_e = \frac{(A_i)(B_j)}{N}$	$\frac{(f_o - f_e)^2}{f_e}$
15	$\frac{40 \times 60}{150} = 16$	0.0625
5	$\frac{40 \times 30}{150} = 8$	1.125
20	$\frac{40 \times 60}{150} = 16$	1
20	$\frac{50 \times 60}{150} = 20$	0
10	$\frac{50 \times 30}{150} = 10$	0
20	$\frac{50 \times 60}{150} = 20$	0
25	$\frac{60 \times 60}{150} = 24$	0.042
15	$\frac{60 \times 30}{150} = 12$	0.75
20	$\frac{60 \times 60}{150} = 24$	0.666
		$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 3.6465$

(v) Critical value: $v = (r - 1)(c - 1) = (3 - 1)(3 - 1) = 4$

$$\chi^2_{0.05}(v = 4) = 9.49$$

(vi) Decision: Since $\chi^2 < \chi^2_{0.05}$, the null hypothesis is accepted at 5% level of significance, i.e., there is no association between two attributes, hair and eye colours.

Example 5

Two researchers adopted different sampling techniques while investigating some group of students to find the number of students falling into different intelligence level. The results are as follows:

Researchers	*Below average*	*Average*	*Above average*	*Genius*	*Total*
X	86	60	44	10	200
Y	40	33	25	2	100
Total	126	93	69	12	300

Would you say that the sampling techniques adopted by the two researchers are significantly different?

Solution

$$N = 300$$

(i) Null Hypothesis H_0: There is no significant difference in the sampling techniques adopted by the two researchers.

(ii) Alternative Hypothesis H_1: There is significant difference in the sampling techniques adopted by the two researchers.

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic:

Observed frequency f_o	*Expected frequency* $f_e = \frac{(A_i)(B_j)}{N}$	$\frac{(f_o - f_e)^2}{f_e}$
86	$\frac{200 \times 126}{300} = 84$	0.0476
60	$\frac{200 \times 93}{300} = 62$	0.0645
44	$\frac{200 \times 69}{300} = 46$	0.0869
10	$\frac{200 \times 12}{300} = 8$	0.5
40	$\frac{100 \times 126}{300} = 42$	0.0952
33	$\frac{100 \times 93}{300} = 31$	0.129
25	$\frac{100 \times 69}{300} = 23$	0.1739
2	$\frac{100 \times 12}{300} = 4$	1
		$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 2.0971$

(v) Critical value: $v = (r-1)(c-1) = (2-1)(4-1) = 3$

$$\chi^2_{0.05}(v = 3) = 7.81$$

(vi) Decision: Since $\chi^2 < \chi^2_{0.05}$, the null hypothesis is accepted at 5% level of significance, i.e., there is no significant difference in the sampling techniques adopted by the two researchers.

Example 6

The following table gives the level of education and the marriage adjustment score for a sample of married women:

Level of education	*Marriage adjustment*				*Total*
	Very low	*Low*	*High*	*Very high*	
College	24	97	62	58	241
High school	22	28	30	41	121
Middle school	32	10	11	20	73
Total	78	135	103	119	435

Can you conclude from the above data the higher the level of education, the greater is the degree of adjustment in marriage?

Solution

$$N = 435$$

(i) Null Hypothesis H_0: There is no relation between the level of education and adjustment in marriage, i.e., two attributes are independent.

(ii) Alternative Hypothesis H_1: There is relation between level of education and adjustment in marriage.

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic:

Observed frequency f_o	Expected frequency $f_e = \frac{(A_i)(B_j)}{N}$	$\frac{(f_o - f_e)^2}{f_e}$
24	$\frac{241 \times 78}{435} \approx 43$	8.3953
97	$\frac{241 \times 135}{435} \approx 75$	6.4533
62	$\frac{241 \times 103}{435} \approx 57$	0.4386
58	$\frac{241 \times 119}{435} \approx 66$	0.9697
22	$\frac{121 \times 78}{435} \approx 22$	0
28	$\frac{121 \times 135}{435} \approx 37$	2.1892
30	$\frac{121 \times 103}{435} \approx 29$	0.0345
41	$\frac{121 \times 119}{435} \approx 33$	1.9394
32	$\frac{73 \times 78}{435} \approx 13$	27.7692
10	$\frac{73 \times 135}{435} \approx 23$	7.3478
11	$\frac{73 \times 103}{435} \approx 17$	2.1176
20	$\frac{73 \times 119}{435} \approx 20$	0
		$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 57.713$

(v) Critical value: $v = (r-1)(c-1) = (3-1)(4-1) = 6$

$$\chi^2_{0.05}\,(v = 6) = 12.59$$

(vi) Decision: Since $\chi^2 > \chi^2_{0.05}$, the null hypothesis is rejected at 5% level of significance i.e., level of education and adjustment in marriage are related and higher the level of education, the greater is the degree of adjustment in marriage.

Example 7

Two batches each of 12 animals are taken for test of inoculation. One batch was inoculated and the other batch was not inoculated. The number of dead and surviving animals are given in the following table for both the cases. Can the inoculation be regarded as effective against the disease. Make Yate's correction for continuity of χ^2?

	Dead	*Survived*	*Total*
Inoculated	2	10	12
Not inoculated	8	4	12
Total	10	14	24

Solution

$$N = 24$$

(i) Null hypothesis H_0: There is no relation between inoculation and death i.e., inoculation and effect on disease are independent.

(ii) Alternative Hypothesis H_1: There is relation between inoculation and death.

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: Yate's correction is used only when $v = 1$ and when some expected frequencies are small, i.e., less than 10. Here, expected frequencies are less than 10 each.

Observed frequency f_o	Expected frequency $f_e = \frac{(A_i)(B_j)}{N}$	$\frac{\{\lvert f_o - f_e\rvert - 0.5\}^2}{f_e}$
2	$\frac{12 \times 10}{25} = 5$	1.25
10	$\frac{12 \times 14}{24} = 7$	0.89
8	$\frac{12 \times 10}{24} = 5$	1.25
4	$\frac{12 \times 14}{24} = 7$	0.89
		$\chi^2 = \sum \frac{\{\lvert f_o - f_e\rvert - 0.5\}^2}{f_e} = 4.28$

(v) Critical value: $v = (2-1)(2-1) = 1$

$$\chi^2_{0.05}\ (v = 1) = 3.84$$

(vi) Decision: Since $\chi^2 > \chi^2_{0.05}$, the null hypothesis is rejected at 5% level of significance, i.e., there is association between inoculation and death and inoculation is regarded as effective against the disease.

EXERCISE 6.5

1. A dice is thrown 264 times with the following results: Show that the dice is biased [Given $\chi^2_{0.05} = 11.07$ for 5 df]

No. appeared on the dice	1	2	3	4	5	6
Frequency	40	32	28	58	54	52

2. A pair of dice are thrown 360 times and frequency of each sum is given below:

Sum	2	3	4	5	6	7	8	9	10	11	12
Frequency	8	24	35	37	44	65	51	42	26	14	14

would you say that the dice are fair on the basis of the chi-square test at 0.05 level of significance?

[Ans.: The dice are fair]

3. 4 coins are tossed 160 times and the following results were obtained:

No. of heads	0	1	2	3	4
Observed frequencies	17	52	54	31	6

Under the assumption that coins are balanced, find the expected frequencies of 0, 1, 2, 3 or 4 heads, and test the goodness of fit ($\alpha = 0.05$).

[Ans.: Expected frequencies: 10, 40, 60, 40, 10, the data do not follow binomial distribution]

4. Fit a Poisson distribution to the following data and for its goodness of fit at level of significance 0.05:

x	0	1	2	3	4
f	419	352	154	56	19

5. The following table gives the number of accidents in a city during a week. Find whether the accidents are uniformly distributed over a week.

Day	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Total
No. of accidents	13	15	9	11	12	10	14	84

[Ans.: The accidents are uniformly distributed over a week]

6. Weights in kilograms of 10 students are given below: 38, 40, 45, 53, 47, 43, 55, 48, 52, 49

Can we say that the variance of the normal distribution from which the above sample is drawn is 20 kg?

[Ans.: The sample is drawn from the normal population with variance 20]

7. Five dice are thrown 192 times and the number of times 4, 5 or 6 are obtained are as follows:

No. of dice showing 4, 5, 6	5	4	3	2	1	0
Frequency	6	46	70	48	20	2

Calculate χ^2.

[Ans.: 16.94]

8. The distribution of defects in printed circuit board is hypothesised to follow Poisson distribution. A random sample of 60 printed boards shows the following data:

No. of defects	0	1	2	3
Observed frequency	32	15	9	4

Does the hypothesis of Poisson distribution appropriate?

[Ans.: The defects follow Poisson distribution]

9. Based on the following data, determine if there is a relation between literacy and smoking.

	Smokers	*Non-smokers*
Literates	83	57
Illiterates	45	68

[Ans.: $\chi^2 = 9.19$, yes]

10. Table below shows the performances of students in mathematics and physics. Test the hypothesis that the performance in mathematics is independent of performance in physics.

Grades in Physics	*Grades in Mathematics*		
	High	*Medium*	*Low*
High	56	71	12
Medium	47	163	38
Low	14	42	81

[Ans.: $\chi^2 = 132.31$, Hypothesis is rejected]

11. Investigate the association between the darkness of eye colour in father and son from the following data:

Colour of son's eyes	*Colour of father's eyes*		*Total*
	Dark	*Not dark*	
Dark	48	90	138
Not dark	80	782	862
Total	128	872	1000

[Ans.: $\chi^2 = 3.84$, There is association between two attributes]

12. From the following data, find whether there is any significant linking in the habit of taking soft drinks among the categories of employees.

Soft drink	*Employees*		
	Clerks	*Teachers*	*Officers*
Pepsi	10	25	65
Thumsup	15	30	65
Fanta	50	60	30

[Ans.: $\chi^2 = 60.24$, Two attributes are not independent]

13. 1000 students at college level were graded according to their IQ and the economic conditions of their home. Use χ^2-test to find out whether there is any association between condition at home and IQ.

Economic condition	*IQ*		
	High	*Low*	*Total*
Rich	460	140	600
Poor	240	160	400
Total	700	300	1000

[Ans.: $\chi^2 = 31.733$, These is no association between two attributes]

14. A random sample of 500 students were classified according to economic condition of their family and also according to merit as shown below:

Merit	*Economic condition*			*Total*
	Rich	*Middleclass*	*Poor*	
Meritorious	42	137	61	240
Not-meritorious	58	113	89	260
Total	100	250	150	500

Test whether the two attributes merit and economic condition are associated or not.

[Ans.: $\chi^2 = 9.30$, The two attributes are associated]

CHAPTER 7

Curve Fitting

Chapter Outline

7.1 Introduction
7.2 Least Square Method
7.3 Fitting of Linear Curves
7.4 Fitting of Quadratic Curves
7.5 Fitting of Exponential and Logarithmic Curves

7.1 INTRODUCTION

Curve fitting is the process of finding the 'best-fit' curve for a given set of data. It is the representation of the relationship between two variables by means of an algebraic equation. On the basis of this mathematical equation, predictions can be made in many statistical problems.

Suppose a set of n points of values (x_1, y_1), (x_2, y_2), …, (x_n, y_n) of the two variables x and y are given. These values are plotted on a rectangular coordinate system, i.e., the xy-plane. The resulting set of points is known as a *scatter diagram* (Fig. 7.1). The scatter diagram exhibits the trend and it is possible to visualize a smooth curve approximating the data. Such a curve is known as an *approximating curve*.

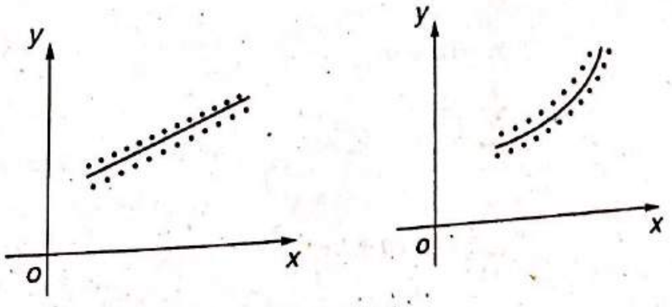


Fig. 7.1

7.2 LEAST SQUARE METHOD

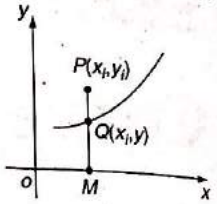


Fig. 7.2

From a scatter diagram, generally, more than one curve may be seen to be appropriate to the given set of data. The method of least squares is used to find a curve which passes through the maximum number of points.

Let P (x_i, y_i) be a point on the scatter diagram (Fig. 7.2). Let the ordinate at P meet the curve $y = f(x)$ at Q and the x-axis at M.

Distance
$$\begin{aligned} QP &= MP - MQ \\ &= y_i - y \\ &= y_i - f(x_i) \end{aligned}$$

The distance QP is known as *deviation*, *error*, or *residual* and is denoted by d_i. It may be positive, negative, or zero depending upon whether P lies above, below, or on the curve. Similar residuals or errors corresponding to the remaining $(n-1)$ points may be obtained. The sum of squares of residuals, denoted by E, is given as

$$E = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} [y_i - f(x_i)]^2$$

If $E = 0$ then all the n points will lie on $y = f(x)$. If $E \neq 0$, $f(x)$ is chosen such that E is minimum, i.e., the best fitting curve to the set of points is that for which E is minimum. This method is known as the least square method. This method does not attempt to determine the form of the curve $y = f(x)$ but it determines the values of the parameters of the equation of the curve.

7.3 FITTING OF LINEAR CURVES

Let (x_i, y_i), $i = 1, 2, \ldots, n$ be the set of n values and let the relation between x and y be $y = a + bx$. The constants a and b are selected such that the straight line is the best fit to the data.

The residual at $x = x_i$ is

$$\begin{aligned} d_i &= y_i - f(x_i) \\ &= y_i - (a + bx_i) \qquad i = 1, 2, \ldots, n \\ E &= \sum_{i=1}^{n} d_i^2 \\ &= \sum_{i=1}^{n} \left[y_i - (a + bx_i)\right]^2 \\ &= \sum_{i=1}^{n} (y_i - a - bx_i)^2 \end{aligned}$$

For E to be minimum,

(i) $\dfrac{\partial E}{\partial a} = 0$

$$\sum_{i=1}^{n} 2(y_i - a - bx_i)(-1) = 0$$

$$\sum_{i=1}^{n} (y_i - a - bx_i) = 0$$

$$\sum_{i=1}^{n} y_i = a\sum_{i=1}^{n} 1 + b\sum_{i=1}^{n} x_i$$

$$\sum y = na + b\sum x$$

(ii) $\dfrac{\partial E}{\partial b} = 0$

$$\sum_{i=1}^{n} 2(y_i - a - bx_i)(-x_i) = 0$$

$$\sum_{i=1}^{n} (x_i y_i - ax_i - bx_i^2) = 0$$

$$\sum_{i=1}^{n} x_i\, y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$$

$$\sum xy = a\sum x + b\sum x^2$$

These two equations are known as *normal equations*. These equations can be solved simultaneously to give the best values of a and b. The best fitting straight line is obtained by substituting the values of a and b in the equation $y = a + bx$.

Example 1

Fit a straight line to the following data:

x	1	2	3	4	6	8
y	2.4	3	3.6	4	5	6

Solution

Let the straight line to be fitted to the data be

$$y = a + bx$$

The normal equations are

$$\sum y = na + b\sum x \qquad \ldots(1)$$

$$\sum xy = a\sum x + b\sum x^2 \qquad \ldots(2)$$

Here, $n = 6$

x	y	x^2	xy
1	2.4	1	2.4
2	3	4	6
3	3.6	9	10.8
4	4	16	16
6	5	36	30
8	6	64	48
$\sum x = 24$	$\sum y = 24$	$\sum x^2 = 130$	$\sum xy = 113.2$

Substituting these values in Eqs (1) and (2),

$$24 = 6a + 24b \quad \text{...(3)}$$
$$113.2 = 24a + 130b \quad \text{...(4)}$$

Solving Eqs (3) and (4),

$$a = 1.9764$$
$$b = 0.5059$$

Hence, the required equation of the straight line is

$$y = 1.9764 + 0.5059x$$

Note $\sum x, \sum y, \sum x^2, \sum xy$ can be directly obtained with the help of scientific calculator.

Example 2

Fit a straight line to the following data. Also, estimate the value of y at $x = 2.5$.

x	0	1	2	3	4
y	1	1.8	3.3	4.5	6.3

Solution

Let the straight line to be fitted to the data be

$$y = a + bx$$

The normal equations are

$$\sum y = na + b\sum x \quad \text{...(1)}$$
$$\sum xy = a\sum x + b\sum x^2 \quad \text{...(2)}$$

Here, $n = 5$

x	y	x^2	xy
0	1	0	0
1	1.8	1	1.8
2	3.3	4	6.6
3	4.5	9	13.5
4	6.3	16	25.2
$\sum x = 10$	$\sum y = 16.9$	$\sum x^2 = 30$	$\sum xy = 47.1$

Substituting these values in Eqs (1) and (2),

$$16.9 = 5a + 10b \quad \text{...(3)}$$
$$47.1 = 10a + 30b \quad \text{...(4)}$$

Solving Eqs (3) and (4),

$$a = 0.72$$
$$b = 1.33$$

Hence, the required equation of the straight line is

$$y = 0.72 + 1.33x$$

At $x = 2.5$,

$$y\,(2.5) = 0.72 + 1.33\,(2.5) = 4.045$$

Example 3

A simply supported beam carries a concentrated load P(lb) at its midpoint. Corresponding to various values of P, the maximum deflection Y(in) is measured. The data is given below:

P	100	120	140	160	180	200
Y	0.45	0.55	0.60	0.70	0.80	0.85

Find a law of the form $Y = a + bP$ using the least square method.

[Summer 2015]

Solution

Let the straight line to be fitted to the data be

$$Y = a + bP$$

The normal equations are

$$\sum Y = na + b\sum P \quad \text{...(1)}$$

$$\sum PY = a\sum P + b\sum P^2 \qquad \text{...(2)}$$

Here, $n = 6$

P	Y	P^2	PY
100	0.45	10000	45
120	0.55	14400	66
140	0.60	19600	84
160	0.70	25600	112
180	0.80	32400	144
200	0.85	40000	170
$\sum P = 900$	$\sum Y = 3.95$	$\sum P^2 = 142000$	$\sum PY = 621$

Substituting these values in Eqs (1) and (2),

$$3.95 = 6a + 900\,b \qquad \text{...(3)}$$

$$621 = 900\,a + 142000\,b \qquad \text{...(4)}$$

Solving Eqs (3) and (4),

$$a = 0.0476$$

$$b = 0.0041$$

Hence, the required equation of the straight line is

$$Y = 0.0476 + 0.0041\,P$$

Example 4

Fit a straight line to the following data. Also, estimate the value of y at $x = 70$.

x	71	68	73	69	67	65	66	67
y	69	72	70	70	68	67	68	64

Solution

Since the values of x and y are larger, we choose the origin for x and y at 69 and 67 respectively,

Let $X = x - 69$ and $Y = y - 67$

Let the straight line to be fitted to the data be

$$Y = a + bX$$

The normal equations are

$$\sum Y = na + b\sum X \qquad \text{...(1)}$$

$$\sum XY = a\sum X + b\sum X^2 \qquad \text{...(2)}$$

Here, $n = 8$

x	y	X	Y	X^2	XY
71	69	2	2	4	4
68	72	−1	5	1	−5
73	70	4	3	16	12
69	70	0	3	0	0
67	68	−2	1	4	−2
65	67	−4	0	16	0
66	68	−3	1	9	−3
67	64	−2	−3	4	6
		$\sum X = -6$	$\sum Y = 12$	$\sum X^2 = 54$	$\sum XY = 12$

Substituting these values in Eqs (1) and (2),

$$12 = 8a - 6b \qquad \text{...(3)}$$

$$12 = -6a + 54b \qquad \text{...(4)}$$

Solving Eqs (3) and (4),

$$a = 1.8182$$

$$b = 0.4242$$

Hence, the required equation of the straight line is

$$Y = 1.8182 + 0.4242X$$

$$y - 67 = 1.8182 + 0.4242(x - 69)$$

$$y = 0.4242x + 39.5484$$

$$y(x = 70) = 0.4242(70) + 39.5484 = 69.2424$$

Note Since $\sum x$, $\sum y$, $\sum x^2$, $\sum xy$ can be directly obtained with the help of scientific calculator, the problem can be solved without shifting the origin.

Example 5

Fit a straight line to the following data taking x as the dependent variable.

x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9

Solution

If x is considered the dependent variable and y the independent variable, the equation of the straight line to be fitted to the data is

$$x = a + by$$

The normal equations are

$$\sum x = na + b\sum y \qquad \dots(1)$$

$$\sum xy = a\sum y + b\sum y^2 \qquad \dots(2)$$

Here, $n = 8$

x	y	y^2	xy
1	1	1	1
3	2	4	6
4	4	16	16
6	4	16	24
8	5	25	40
9	7	49	63
11	8	64	88
14	9	81	126
$\sum x = 56$	$\sum y = 40$	$\sum y^2 = 256$	$\sum xy = 364$

Substituting these values in Eqs (1) and (2),

$$56 = 8a + 40b \qquad \dots(3)$$

$$364 = 40a + 256b \qquad \dots(4)$$

Solving Eqs (3) and (4),

$$a = -0.5$$

$$b = 1.5$$

Hence, the required equation of the straight line is

$$x = -0.5 + 1.5y$$

Example 6

If P is the pull required to lift a load W by means of a pulley block, find a linear law of the form P = mW + c connecting P and W using the following data:

P	12	15	21	25
W	50	70	100	120

where P and W are taken in kg-wt. Compute P when W = 150 kg.

Solution

Let the linear curve (straight line) fitted to the data be

$$P = mW + c = c + mW$$

The normal equations are

$$\sum P = nc + m\sum W \qquad \dots(1)$$

$$\sum PW = c\sum W + m\sum W^2 \qquad \dots(2)$$

Here, $n = 4$

P	W	W^2	PW
12	50	2500	600
15	70	4900	1050
21	100	10000	2100
25	120	14400	3000
$\sum P = 73$	$\sum W = 340$	$\sum W^2 = 31800$	$\sum PW = 6750$

Substituting these values in Eqs (1) and (2),

$$73 = 4c + 340\,m \qquad \dots(3)$$

$$6750 = 340\,c + 31800\,m \qquad \dots(4)$$

Solving Eqs (3) and (4),

$$c = 2.2759$$

$$m = 0.1879$$

Hence, the required equation of the straight line is

$$P = 0.1879\,W + 2.2759$$

When $W = 150$ kg,

$$P = 0.1879(150) + 2.2759 = 30.4609$$

EXERCISE 7.1

1. Fit the line of best fit to the following data:

x	0	5	10	15	20	25
y	12	15	17	22	24	30

[Ans.: $y = 0.7x + 11.28$]

2. The results of a measurement of electric resistance R of a copper bar at various temperatures t°C are listed below:

t°C	19	25	30	36	40	45	50
R	76	77	79	80	82	83	85

Find a relation $R = a + bt$ where a and b are constants to be determined.

[Ans.: $R = 70.0534 + 0.2924t$]

3. Fit a straight line to the following data:

x	1.53	1.78	2.60	2.95	3.42
y	33.50	36.30	40.00	45.85	53.40

[**Ans.:** $y = 19 + 9.7x$]

4. Fit a straight line to the following data:

x	100	120	140	160	180	200
y	0.45	0.55	0.60	0.70	0.80	0.85

[**Ans.:** $y = 0.0475 + 0.00407x$]

5. Find the relation of the type $R = aV + b$, when some values of R and V obtained from an experiment are

V	60	65	70	75	80	85	90
R	109	114	118	123	127	130	133

[**Ans.:** $R = 0.8071V + 61.4675$]

7.4 FITTING OF QUADRATIC CURVES

Let (x_i, y_i), $i = 1, 2, \ldots, n$ be the set of n values and let the relation between x and y be $y = a + bx + cx^2$. The constants a, b, and c are selected such that the parabola is the best fit to the data. The residual at $x = x_i$ is

$$\begin{aligned} d_i &= y_i - f(x_i) \\ &= y_i - \left(a + bx_i + cx_i^2\right) \end{aligned}$$

$$\begin{aligned} E &= \sum_{i=1}^{n} d_i^2 \\ &= \sum_{i=1}^{n}\left[y_i - \left(a + bx_i + cx_i^2\right)\right]^2 \\ &= \sum_{i=1}^{n}\left(y_i - a - bx_i - cx_i^2\right)^2 \end{aligned}$$

For E to be minimum,

(i) $\dfrac{\partial E}{\partial a} = 0$

$$\sum_{i=1}^{n} 2\left(y_i - a - bx_i - cx_i\right)(-1) = 0$$

$$\sum_{i=1}^{n}\left(y_i - a - bx_i - cx_i\right) = 0$$

$$\sum_{i=1}^{n} y_i = a\sum_{i=1}^{n} 1 + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2$$

$$\sum y = na + b\sum x + c\sum x^2$$

(ii) $\dfrac{\partial E}{\partial b} = 0$

$$\sum_{i=1}^{n} 2(y_i - a - bx_i - cx_i)(-x_i) = 0$$

$$\sum_{i=1}^{n}\left(x_i y_i - ax_i - bx_i^2 - cx_i^3\right) = 0$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3$$

$$\sum xy = na + b\sum x^2 + c\sum x^3$$

(iii) $\dfrac{\partial E}{\partial c} = 0$

$$\sum_{i=1}^{n} 2(y_i - a - bx_i - cx_i^2)(x_i^2) = 0$$

$$\sum_{i=1}^{n} x_i^2 y_i - ax_i^2 - bx_i^3 - cx_i^4 = 0$$

$$\sum_{i=1}^{n} x_i^2 y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4$$

$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4$$

These equations are known as *normal equations*. These equations can be solved simultaneously to give the best values of a, b, and c. The best fitting parabola is obtained by substituting the values of a, b, and c in the equation $y = a + bx + cx^2$.

Example 1

Fit a least squares quadratic curve to the following data:

x	1	2	3	4
y	1.7	1.8	2.3	3.2

Estimate $y(2.4)$.

Solution

Let the equation of the least squares quadratic curve (parabola) be $y = a + bx + cx^2$.
The normal equations are

$$\sum y = na + b\sum x + c\sum x^2 \quad \ldots(1)$$
$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3 \quad \ldots(2)$$
$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4 \quad \ldots(3)$$

Here, $n = 4$

x	y	x^2	x^3	x^4	xy	x^2y
1	1.7	1	1	1	1.7	1.7
2	1.8	4	8	16	3.6	7.2
3	2.3	9	27	81	6.9	20.7
4	3.2	16	64	256	12.8	51.2
$\Sigma x = 10$	$\Sigma y = 9$	$\Sigma x^2 = 30$	$\Sigma x^3 = 100$	$\Sigma x^4 = 354$	$\Sigma xy = 25$	$\Sigma x^2y = 80.8$

Substituting these values in Eqs (1), (2), and (3),

$$9 = 4a + 10b + 30c \quad \ldots(4)$$
$$25 = 10a + 30b + 100c \quad \ldots(5)$$
$$80.8 = 30a + 100b + 354c \quad \ldots(6)$$

Solving Eqs (4), (5), and (6),

$$a = 2$$
$$b = -0.5$$
$$c = 0.2$$

Hence, the required equation of least squares quadratic curve is

$$y = 2 - 0\cdot 5x + 0\cdot 2x^2$$
$$y(2\cdot 4) = 2 - 0\cdot 5(2\cdot 4) + 0\cdot 2(2.4)^2 = 1\cdot 952$$

Note Σx, Σy, Σx^2, Σx^3, Σx^4, Σxy, Σx^2y can be directly obtained with the help of scientific calculator.

Example 2

Fit a second-degree polynomial using least square method to the following data:

x	0	1	2	3	4
y	1	1.8	1.3	2.5	6.3

[Summer 2015]

Solution

Let the equation of the least squares quadratic curve be $y = a + bx + cx^2$. The normal equations are

$$\sum y = na + b\sum x + c\sum x^2 \quad \ldots(1)$$
$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3 \quad \ldots(2)$$
$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4 \quad \ldots(3)$$

Here, $n = 5$

x	y	x^2	x^3	x^4	xy	x^2y
0	1	0	0	0	0	0
1	1.8	1	1	1	1.8	1.8
2	1.3	4	8	16	2.6	5.2
3	2.5	9	27	81	7.5	22.5
4	6.3	16	64	256	25.2	100.8
$\Sigma x = 10$	$\Sigma y = 12.9$	$\Sigma x^2 = 30$	$\Sigma x^3 = 100$	$\Sigma x^4 = 354$	$\Sigma xy = 37.1$	$\Sigma x^2y = 130.3$

Substituting these values in Eqs (1), (2), and (3),

$$12.9 = 5a + 10b + 30c \quad \ldots(4)$$
$$37.1 = 10a + 30b + 100c \quad \ldots(5)$$
$$130.3 = 30a + 100b + 354c \quad \ldots(6)$$

Solving Eqs (4), (5), and (6),

$$a = 1.42$$
$$b = -1.07$$
$$c = 0.55$$

Hence, the required equation of the least squares quadratic curve is

$$y = 1.42 - 1.07x + 0.55x^2$$

Example 3

By the method of least squares, fit a parabola to the following data:

x	1	2	3	4	5
y	5	12	26	60	97

Also, estimate y at x = 6.

Solution

Let the equation of the parabola be $y = a + bx + cx^2$. The normal equations are

$$\sum y = na + b\sum x + c\sum x^2 \quad \ldots(1)$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3 \quad \text{...(2)}$$

$$\sum x^2y = a\sum x^2 + b\sum x^3 + c\sum x^4 \quad \text{...(3)}$$

Here, $n = 5$

x	y	x^2	x^3	x^4	xy	x^2y
1	5	1	1	1	5	5
2	12	4	8	16	24	48
3	26	9	27	81	78	234
4	60	16	64	256	240	960
5	97	25	125	625	485	2425
$\sum x = 15$	$\sum y = 200$	$\sum x^2 = 55$	$\sum x^3 = 225$	$\sum x^4 = 979$	$\sum xy = 832$	$\sum x^2y = 3672$

Substituting these values in Eqs (1), (2), and (3),

$$200 = 5a + 15b + 55c \quad \text{...(4)}$$

$$832 = 15a + 55b + 225c \quad \text{...(5)}$$

$$3672 = 55a + 225b + 979c \quad \text{...(6)}$$

Solving Eqs (4), (5), and (6),

$$a = 10.4$$

$$b = -11.0857$$

$$c = 5.7143$$

Hence, the required equation of the parabola is

$$y = 10.4 - 11.0857\,x + 5.7143\,x^2$$

$$y(6) = 10.4 - 11.0857(6) + 5.7143(6)^2 = 149.6006$$

Example 4

Fit a second-degree parabolic curve to the following data.

x	1	2	3	4	5	6	7	8	9
y	2	6	7	8	10	11	11	10	9

Solution

Let

$$X = x - 5$$

$$Y = y - 10$$

Let the equation of the parabola be $Y = a + bX + cX^2$.

The normal equations are

$$\sum Y = na + b\sum X + c\sum X^2 \quad \text{...(1)}$$

$$\sum XY = a\sum X + b\sum X^2 + c\sum X^3 \quad \text{...(2)}$$

$$\sum X^2Y = a\sum X^2 + b\sum X^3 + c\sum X^4 \quad \text{...(3)}$$

Here, $n = 9$

x	y	X	Y	X^2	X^3	X^4	XY	X^2Y
1	2	−4	−8	16	−64	256	32	−128
2	6	−3	−4	9	−27	81	12	−36
3	7	−2	−3	4	−8	16	6	−12
4	8	−1	−2	1	−1	1	2	−2
5	10	0	0	0	0	0	0	0
6	11	1	1	1	1	1	1	1
7	11	2	1	4	8	16	2	4
8	10	3	0	9	27	81	0	0
9	9	4	−1	16	64	256	−4	−16
		$\Sigma X = 0$	$\Sigma Y = -16$	$\Sigma X^2 = 60$	$\Sigma X^3 = 0$	$\Sigma X^4 = 708$	$\Sigma XY = 51$	$\Sigma X^2Y = -189$

Substituting these values in Eqs (1), (2), and (3),

$$-16 = 9a + 60c \quad \text{...(4)}$$

$$51 = 60b \quad \text{...(5)}$$

$$-189 = 60a + 708c \quad \text{...(6)}$$

Solving Eqs (4), (5), and (6),

$$a = 0.0043$$

$$b = 0.85$$

$$c = -0.2673$$

Hence, the required equation of the parabola is

$$Y = 0.0043 + 0.85X - 0.2673X^2$$

$$y - 10 = 0.0043 + 0.85(x-5) - 0.2673(x-5)^2$$

$$\begin{aligned} y &= 10 + 0.0043 + 0.85(x-5) - 0.2673(x^2 - 10x + 25) \\ &= 10 + 0.0043 + 0.85x - 4.25 - 0.2673x^2 + 2.673x - 6.6825 \\ &= -0.9282 + 3.523x - 0.2673x^2 \end{aligned}$$

Note Since $\sum x, \sum y, \sum x^2, \sum x^3, \sum x^4, \sum xy, \sum x^2y$ can be directly obtained with the help of scientific calculator, the problem can be solved without shifting the origin.

Example 5

Fit a second-degree parabola $y = a + bx^2$ to the following data:

x	1	2	3	4	5
y	1.8	5.1	8.9	14.1	19.8

Solution

Let the curve to be fitted to the data be

$$y = a + bx^2$$

The normal equations are

$$\sum y = na + b\sum x^2 \quad \text{...(1)}$$

$$\sum x^2 y = a\sum x^2 + b\sum x^4 \quad \text{...(2)}$$

Here, $n = 5$

x	y	x^2	x^4	x^2y
1	1.8	1	1	1.8
2	5.1	4	16	20.4
3	8.9	9	81	80.1
4	14.1	16	256	225.6
5	19.8	25	625	495
	$\Sigma y = 49.7$	$\Sigma x^2 = 55$	$\Sigma x^4 = 979$	$\Sigma x^2y = 822.9$

Substituting these values in Eqs (1) and (2),

$$49.7 = 5a + 55b \quad \text{...(3)}$$

$$822.9 = 55a + 979\,b \quad \text{...(4)}$$

Solving Eqs (3) and (4),

$$a = 1.8165$$

$$b = 0.7385$$

Hence, the required equation of the curve is

$$y = 1.8165 + 0.7385\,x^2$$

Example 6

Fit a curve $y = ax + bx^2$ *for the following data:*

x	1	2	3	4	5	6
y	2.51	5.82	9.93	14.84	20.55	27.06

Solution

Let the curve to be fitted to the data be

$$y = ax + bx^2$$

The normal equations are

$$\sum xy = a\sum x^2 + b\sum x^3 \quad \text{...(1)}$$

$$\sum x^2 y = a\sum x^3 + b\sum x^4 \quad \text{...(2)}$$

x	y	x^2	x^3	x^4	xy	x^2y
1	2.51	1	1	1	2.51	2.51
2	5.82	4	8	16	11.64	23.28
3	9.93	9	27	81	29.79	89.37
4	14.84	16	64	256	59.36	237.44
5	20.55	25	125	625	102.75	513.75
6	27.06	36	216	1296	162.36	974.16
		$\Sigma x^2 = 91$	$\Sigma x^3 = 441$	$\Sigma x^4 = 2275$	$\Sigma xy = 368.41$	$\Sigma x^2y = 1840.51$

Substituting these values in Eqs (1) and (2),

$$368 \cdot 41 = 91a + 441b \quad \text{...(3)}$$

$$1840 \cdot 51 = 441\,a + 2275\,b \quad \text{...(4)}$$

Solving Eqs (3) and (4),

$$a = 2.11$$

$$b = 0.4$$

Hence, the required equation of the curve is

$$y = 2 \cdot 11x + 0 \cdot 4x^2$$

EXERCISE 7.2

1. Fit a parabola to the following data:

x	–2	–1	0	1	2
y	1.0	1.8	1.3	2.5	6.3

[Ans. : $y = 1.48 + 1.13x + 0.55x^2$]

2. Fit a curve $y = ax + bx^2$ to the following data:

x	–2	–1	0	1	2
y	–72	–46	–12	35	93

[Ans. : $y = 41.1x + 2.147x^2$]

3. Fit a parabola $y = a + bx + cx^2$ to the following data:

x	0	2	5	10
y	4	7	6.4	−6

[Ans.: $y = 4.1 + 1.979x - 0.299x^2$]

4. Fit a curve $y = a_0 + a_1x + a_2x^2$ for the given data:

x	3	5	7	9	11	13
y	2	3	4	6	5	8

[Ans.: $y = 0.7897 + 0.4004x + 0.0089x^2$]

7.5 FITTING OF EXPONENTIAL AND LOGARITHMIC CURVES

Let (x_i, y_i), $i = 1, 2, \ldots, n$ be the set of n values and let the relation between x and y be $y = ab^x$.

Taking logarithm on both the sides of the equation $y = ab^x$,

$$\log_e y = \log_e a + x \log_e b$$

Putting $\log_e y = Y$, $\log_e a = A$, $x = X$, and $\log_e b = B$,

$$Y = A + BX$$

This is a linear equation in X and Y. The normal equations are

$$\sum Y = nA + B\sum X$$

$$\sum XY = A\sum X + B\sum X^2$$

Solving these equations, A and B, and, hence, a and b can be found. The best fitting exponential curve is obtained by substituting the values of a and b in the equation $y = ab^x$.

Similarly, the best fitting exponential curves for the relation $y = ax^b$ and $y = ae^{bx}$ can be obtained.

Example 1

Find the law of the form $y = ab^x$ *to the following data:*

x	1	2	3	4	5	6	7	8
y	1	1.2	1.8	2.5	3.6	4.7	6.6	9.1

Solution

$$y = ab^x$$

Taking logarithm on both the sides,

$$\log_e y = \log_e a + x \log_e b$$

Putting $\log_e y = Y$, $\log_e a = A$, $x = X$ and $\log_e b = B$,

$$Y = A + BX$$

The normal equations are

$$\sum Y = nA + B\sum X \quad \ldots(1)$$

$$\sum XY = A\sum X + B\sum X^2 \quad \ldots(2)$$

Here, $n = 8$

x	y	X	Y	X²	XY
1	1	1	0.0000	1	0.0000
2	1.2	2	0.1823	4	0.3646
3	1.8	3	0.5878	9	1.7634
4	2.5	4	0.9163	16	3.6652
5	3.6	5	1.2809	25	6.4045
6	4.7	6	1.5476	36	9.2856
7	6.6	7	1.8871	49	13.2097
8	9.1	8	2.2083	64	17.6664
		$\sum X = 36$	$\sum Y = 8.6103$	$\sum X^2 = 204$	$\sum XY = 52.3594$

Substituting these values in Eqs (1) and (2),

$$8.6103 = 8A + 36B \quad \ldots(3)$$

$$52.3594 = 36A + 204B \quad \ldots(4)$$

Solving Eqs (3) and (4),

$$A = -0.3823$$

$$B = 0.3241$$

$$\log_e a = A$$

$$\log_e a = -0.3823$$

$$a = 0.6823$$

$$\log_e b = B$$

$$\log_e b = 0.3241$$

$$b = 1.3828$$

Hence, the required law is

$$y = 0.6823\,(1.3828)^x$$

Example 2

Fit a curve of the form $y = ab^x$ to the following data by the method of least squares:

x	1	2	3	4	5	6	7
y	87	97	113	129	202	195	193

Solution

$$y = ab^x$$

Taking logarithm on both the sides,

$$\log_e y = \log_e a + x \log_e b$$

Putting $\log_e y = Y$, $\log_e a = A$, $x = X$ and $\log_e b = B$,

$$Y = A + BX$$

The normal equations are

$$\sum Y = nA + B\sum X \qquad \text{...(1)}$$

$$\sum XY = A\sum X + B\sum X^2 \qquad \text{...(2)}$$

Here, $n = 7$

x	y	X	Y	X^2	XY
1	87	1	4.4659	1	4.4659
2	97	2	4.5747	4	9.1494
3	113	3	4.7274	9	14.1822
4	129	4	4.8598	16	19.4392
5	202	5	5.3083	25	26.5415
6	195	6	5.2730	36	31.6380
7	193	7	5.2627	49	36.8389
		$\sum X = 28$	$\sum Y = 34.4718$	$\sum X^2 = 140$	$\sum XY = 142.2551$

Substituting these values in Eqs (1) and (2),

$$34.4718 = 7A + 28\,B \qquad \text{...(3)}$$

$$142.2551 = 28\,A + 140\,B \qquad \text{...(4)}$$

Solving Eqs (3) and (4),

$$A = 4.3006$$

$$B = 0.156$$

$$\log_e a = A$$

$$\log_e a = 4.3006$$

$$a = 73.744$$

$$\log_e b = B$$

$$\log_e b = 0.156$$

$$b = 1.1688$$

Hence, the required curve is

$$y = 73.744\,(1.1688)^x$$

Example 3

Fit a curve of the form $y = ax^b$ to the following data:

x	20	16	10	11	14
y	22	41	120	89	56

Solution

$$y = ax^b$$

Taking logarithm on both the sides,

$$\log_e y = \log_e a + b \log_e x$$

Putting $\log_e y = Y$, $\log_e a = A$, $b = B$ and $\log_e x = X$,

$$Y = A + BX$$

The normal equations are

$$\sum Y = nA + B\sum X \qquad \text{...(1)}$$

$$\sum XY = A\sum X + B\sum X^2 \qquad \text{...(2)}$$

Here, $n = 5$

x	y	X	Y	X^2	XY
20	22	2.9957	3.0910	8.9742	9.2597
16	41	2.7726	3.7136	7.6873	10.2963
10	120	2.3026	4.7875	5.3019	11.0237
11	89	2.3979	4.4886	5.7499	10.7632
14	56	2.6391	4.0254	6.9648	10.6234
		$\sum X = 13.1079$	$\sum Y = 20.1061$	$\sum X^2 = 34.6781$	$\sum XY = 51.9663$

Substituting these values in Eqs (1) and (2),

$$20.1061 = 5A + 13.1079\,B \qquad \text{...(3)}$$

$$51.9663 = 13.1079\,A + 34.6781\,B \qquad \text{...(4)}$$

Solving Eqs (3) and (4),

$$A = 10.2146$$
$$B = -2.3624$$
$$\log_e a = A$$
$$\log_e a = 10.2146$$
$$a = 27298 \cdot 8539$$
$$\text{and } b = B = -2.3624$$

Hence, the required equation of the curve is

$$y = 27298.8539\,x^{-2.3624}$$

Example 4

Fit a curve of the form $y = ae^{bx}$ *to the following data:*

x	1	3	5	7	9
y	115	105	95	85	80

Solution

$$y = ae^{bx}$$

Taking logarithm on both the sides,

$$\log_e y = \log_e a + bx\log_e e$$
$$= \log_e a + bx$$

Putting $\log_e y = Y$, $\log_e a = A$, $b = B$ and $x = X$,

$$Y = A + BX$$

The normal equations are

$$\sum Y = nA + B\sum X \quad \text{...(1)}$$
$$\sum XY = A\sum X + B\sum X^2 \quad \text{...(2)}$$

Here, $n = 5$

x	y	X	Y	X^2	XY
1	115	1	4.7449	1	4.7449
3	105	3	4.6539	9	13.9617
5	95	5	4.5539	25	22.7695
7	85	7	4.4427	49	31.0989
9	80	9	4.3820	81	39.438
		$\sum X = 25$	$\sum Y = 22.7774$	$\sum X^2 = 165$	$\sum XY = 112.013$

Substituting these values in Eqs (1) and (2),

$$22.7774 = 5A + 25B \quad \text{...(3)}$$
$$112.013 = 25A + 165B \quad \text{...(4)}$$

Solving Eqs (3) and (4),

$$A = 4.7897$$
$$B = -0.0469$$
$$\log_e a = A$$
$$\log_e a = 4.7897$$
$$a = 120.2653$$
$$b = B = -0.0469$$

and

Hence, the required equation of the curve is

$$y = 120.2653\,e^{-0.0469x}$$

Example 5

Fit the exponential curve $y = ae^{bx}$ *to the following data:*

x	0	2	4	6	8
y	150	63	28	12	5.6

[Summer 2015]

Solution

$$y = ae^{bx}$$

Taking logarithm on both the sides,

$$\log_e y = \log_e a + bx\log_e e$$
$$= \log_e a + bx$$

Putting $\log_e y = Y$, $\log_e a = A$, $b = B$ and $x = X$,

$$Y = A + BX$$

The normal equations are

$$\sum Y = nA + b\sum X \quad \text{...(1)}$$
$$\sum XY = A\sum X + B\sum X^2 \quad \text{...(2)}$$

Here, $n = 5$

x	y	X	Y	X^2	XY
0	150	0	5.0106	0	0
2	63	2	4.1431	4	8.2862
4	28	4	3.3322	16	13.3288
6	12	6	2.4849	36	14.9094
8	5.6	8	1.7228	64	13.7824
		$\sum X = 20$	$\sum Y = 16.6936$	$\sum X^2 = 120$	$\sum XY = 50.3068$

Substituting these values in Eqs (1) and (2),

$$16.6936 = 5A + 20B \qquad \ldots(3)$$

$$50.3068 = 20A + 120B \qquad \ldots(4)$$

Solving Eqs (3) and (4),

$$A = 4.9855$$

$$B = -0.4117$$

$$\log_e a = A$$

$$\log_e a = 4.9855$$

$$a = 146.28$$

and

$$b = B = -0.4117$$

Hence, the required equation of the curve is

$$y = 146.28\, e^{-0.4117x}$$

Example 6

The pressure and volume of a gas are related by the equation $PV^\gamma = c$. Fit this curve to the following data:

P	0.5	1.0	1.5	2.0	2.5	3.0
V	1.62	1.00	0.75	0.62	0.52	0.46

Solution

$$PV^\gamma = c$$

Taking logarithm on both the sides,

$$\log_e P + \gamma \log_e V = \log_e c$$

$$\log_e V = \frac{1}{\gamma}\log_e c - \frac{1}{\gamma}\log_e P$$

Putting $\log_e V = y, \frac{1}{\gamma}\log_e c = a, \log_e P = x, -\frac{1}{\gamma} = b,$

$$y = a + bx$$

The normal equations are

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

Here, $n = 6$

P	V	x	y	x^2	xy
0.5	1.62	−0.6931	0.4824	0.4804	−0.3343
1.0	1.00	0	0	0	0
1.5	0.75	0.4055	−0.2877	0.1644	−0.1166
2.0	0.62	0.6931	−0.4780	0.4804	−0.3313
2.5	0.52	0.9163	−0.6539	0.8396	−0.5992
3.0	0.46	1.0986	−0.7765	1.2069	−0.8531
		$\sum x = 2.4204$	$\sum y = -1.7137$	$\sum x^2 = 3.1717$	$\sum xy = -2.2345$

Substituting these values in Eqs (1) and (2),

$$-1.7137 = 6a + 2.4204\,b \qquad \ldots(3)$$

$$-2.2345 = 2.4204a + 3.1717\,b \qquad \ldots(4)$$

Solving Eqs (3) and (4),

$$a = -0.002$$

$$b = -0.7029$$

$$-\frac{1}{\gamma} = b$$

$$\gamma = 1.4227$$

$$\frac{1}{\gamma}\log_e c = a$$

$$\frac{1}{1.4227}\log_e c = -0.002$$

$$c = 0.9972$$

Hence, the required equation of the curve is

$$PV^{(1.4227)} = 0.9972$$

EXERCISE 7.3

1. Fit the curve $y = ab^x$ to the following data:

x	2	3	4	5	6
y	144	172.3	207.4	248.8	298.5

[Ans.: $y = 100\,(1.2)$

2. Fit the curve $y = ae^{bx}$ to the following data:

x	0	2	4
y	5.012	10	31.62

[Ans.: $y = 4.642e^{0.46x}$]

3. Fit the curve $y = ax^b$ to the following data:

x	1	2	3	4
y	2.50	8.00	19.00	50.00

[Ans.: $y = 2.227x^{2.09}$]

4. Estimate γ by fitting the ideal gas law $PV^{\gamma} = c$ to the following data:

P	16.6	39.7	78.5	115.5	195.3	546.1
V	50	30	20	15	10	5

[Ans.: $\gamma = 1.504$]